A CLASS OF IPPS SAMPLING SCHEMES

L.N. Sahoo[1#] ; S.C. Senapati**[2]; A.K. Mangaraj***[3]
*Department of Statistics, Utkal University, Bhubaneswar 751004, India
**Department of Statistics, Ravenshaw College, Cuttack 753003, India
***Department of Statistics, R.D. Womens College, Bhubaneswar 751004, India

ABSTRACT
This paper introduces a general class of Inclusion Probability Proportional to Size (IPPS) sampling schemes for selecting two units from a finite population. All IPPS sampling schemes, identified as particular members of this class, possess some desirable properties with regard to the inclusion probabilities, and provide unbiased and non-negative variance estimators under Horvitz-Thomson (HT) model.

KEYWORDS : Inclusion probability, joint inclusion probability, unequal probability sampling.

MSC: 62D05

RESUMEN
Este trabajo introduce una clase general de esquemas de muestreo de Probabilidades de Inclusión Proporcionales al tamaño (IPPS) para seleccionar dos unidades de una población finita. Todos los esquemas de muestreo IPPS, identificados como miembros particulares de esta clase, posee algunas propiedades deseables respecto a las probabilidades de inclusión, y provee estimadores insesgados no negativos de la varianza bajo el modelo de Horvitz-Thomson (HT) .

1. INTRODUCTION

Let y_i be the value of the study variable y, on the *i*th unit of a finite population, $i = 1,2,\ldots,N$. To estimate $Y = \sum_{i=1}^{N} y_i$, the population total of $y-$values, assume that a sample s of n distinct units is selected from the population according to some unequal probability sampling without replacement scheme with π_i as the inclusion probability of *i*th unit and π_{ij} as the joint inclusion probability of *i*th and *j*th units. The most commonly used estimator in this context is the Horvitz and Thompson (1952) (HT) estimator defined by

$$t_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

with variance

$$Var(t_{HT}) = \frac{1}{2} \sum_{i \neq j = 1}^{N} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

[see, for example Sarndal *et al*. (2003)].

Sen (1953), and Yates and Grundy (1953) independently suggested an unbiased estimator of the $Var(t_{HT})$ given by

[1 #] Corresponding author, lnsahoostatuu@rediffmail.com
[2] scsenapati2002@rediffmail.com
[3] akmangaraj@gmail.com

$$v(t_{HT}) = \frac{1}{2}\sum_{i \neq j \in s} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 . \tag{1}$$

A sufficient condition for this expression to be always non-negative is that $\pi_i \pi_j > \pi_{ij}$, $i \neq j$.

It is well known that, considerable reduction in the variance of t_{HT} can be expected if the sampling scheme ensures that π_i 's are proportional to y_i or nearly so. But, in the absence of knowledge on $y-$values there is no scope to investigate such a relationship at the sampling stage. So, in the face of this disadvantage, the sampling schemes which ensure $\pi_i \propto x_i$ are usually employed in practice, where x_i is the value of an auxiliary variable x (supposed to be strongly related to y) for the *i*th unit of the population. Such schemes are termed as Inclusion Probability Proportional to Size (IPPS or πps) sampling schemes. The estimator commonly used with such schemes is the HT estimator. Hence, from the general theory developed by Horvitz and Thompson (1952), an IPPS sampling scheme must satisfy $\sum_{i=1}^{N} \pi_i = n$, $\sum_{i \neq j}^{N} \pi_{ij} = (n-1)\pi_i$ and $\sum_i \sum_{j<i} \pi_{ij} = \frac{1}{2} n(n-1)$. These are known as πps properties of the scheme.

A number of IPPS sampling schemes are available in the literature [*cf*, Brewer and Hanif (1983), and Chaudhuri and Vos (1988)] showing that the development of such schemes has become an endless effort of the survey samplers. There are perhaps two important reasons to take up such an incentive *viz*., (i) a researcher can discover various possible ways for achieving $\pi_i \propto x_i$, (ii) no IPPS sampling scheme is uniformly better than others in all respects. A majority of these methods are restricted to $n=2$ only. Because, calculation of π_{ij} becomes cumbersome for $n>2$ and some procedures seem to be less precise than even probability proportional to size with replacement (PPSWR) scheme. However, an IPPS scheme with $n=2$ is very much useful in stratified sampling, where stratification is sufficiently '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 [*cf*, Chaudhuri and Vos (1988, p.148)].

The purpose of our paper is to develop a general class of IPPS sampling schemes achieving πps requirement and providing an unbiased and non-negative Sen-Yates-Grundy estimator of $Var(t_{HT})$. Although the scheme can be applicable for $n>2$, we are confined to $n=2$ only in order to avoid complexity in deriving expression for π_{ij} .

2. DESCRIPTION OF THE CLASS OF SAMPLING SCHEMES

Assuming that x_i 's are known and positive for all i , let us define $p_i = x_i \Big/ \sum_{i=1}^{N} x_i$ as the initial probability of selection of *i*th unit. Then, corresponding to the set of initial probabilities $\{p_1, p_2, ..., p_N\}$ for the N population units, consider the set of revised probabilities $\{P_1, P_2, ..., P_N\}$, where P_i is defined by

$$P_i = \frac{p_i(1-z_i)(2-\gamma p_i^{\delta-1})}{1-2z_i} \ , \ i = 1,2,\ldots,N,$$

such that $z_i = p_i^{\delta} \Big/ \sum_{i=1}^{N} p_i^{\delta}$, δ being a known constant and $\gamma = \sum_{i=1}^{N} \frac{p_i}{1-2z_i} \Big/ \sum_{i=1}^{N} \frac{p_i^{\delta}(1-z_i)}{1-2z_i}$,

determined so as to make $\sum_{i=1}^{N} P_i = 1$, *i.e*., by solving the equation

$$\sum_{i=1}^{N} \frac{p_i \{1+(1-2z_i)\}}{1-2z_i} - \gamma \sum_{i=1}^{N} \frac{p_i^{\delta}(1-z_i)}{1-2z_i} = 1, \tag{2}$$

for γ.

It must be noted here that the computation of revised probabilities is restricted only to those situations for which $p_i^{\delta-1} \le 2/\gamma$ and $z_i \le \frac{1}{2}\ \forall i$, because otherwise (1) would give negative results.

The suggested class of sampling schemes (S_{δ} , say) consists of the following steps :

Step I. Draw the first unit, say i , with revised probability P_i and without replacement

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

$$P_{j/i} = \frac{z_j}{1-z_i} .$$

3. INCLUSION PROBABILITIES AND PROPERTIES OF S_{δ}

By definition,

$$\pi_i = P_i + \sum_{j \ne i} P_j \frac{z_i}{1-z_j}$$

$$= 2p_i - \gamma p_i^{\delta} + z_i \sum_{i=1}^{N} \frac{p_i(2-\gamma p_i^{\delta-1})}{1-2z_i} . \tag{3}$$

Again, from (2), we have

$$\sum_{i=1}^{N} \frac{p_i}{1-2z_i} - \frac{\gamma}{2} \sum_{i=1}^{N} \frac{p_i^{\delta} \{1+(1-2z_i)\}}{1-2z_i} = 0$$

i.e.,

$$z_i \sum_{i=1}^{N} \frac{p_i(2-\gamma p_i^{\delta-1})}{1-2z_i} - \gamma p_i^{\delta} = 0 . \tag{4}$$

Hence, from (3) and (4) we obtain

$$\pi_i = 2p_i . \tag{5}$$

The second order inclusion probabilities are

$$\pi_{ij} = P_i P_{j/i} + P_j P_{i/j}$$

$$= \frac{p_i z_j (2-\gamma p_i^{\delta-1})}{1-2z_i} + \frac{p_j z_i (2-\gamma p_j^{\delta-1})}{1-2z_j}. \tag{6}$$

The desirable properties of the scheme are as follows:

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

(ii) $\sum_{j\neq i}^{N} \pi_{ij} = \frac{p_i(2-\gamma p_i^{\delta-1})}{1-2z_i} \sum_{j\neq i}^{N} z_j + z_i \sum_{j\neq i}^{N} \frac{p_j(2-\gamma p_j^{\delta-1})}{1-2z_j}$

$$= 2p_i - \gamma p_i^{\delta} + z_i \sum_{j=1}^{N} \frac{p_j(2-\gamma p_j^{\delta-1})}{1-2z_j}$$

$= 2p_i$ [using (4)]

$= \pi_i.$

(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), we obtain

$$\pi_i \pi_j - \pi_{ij} = \frac{p_i p_j (2-\gamma p_i^{\delta-1})(2-\gamma p_j^{\delta-1})}{(1-2z_i)(1-2z_j)} \left(\sum_{k>2} z_k \right)^2 + z_i z_j \left[\sum_{k>2} \frac{p_k(2-\gamma p_k^{\delta-1})}{1-2z_k} \right]^2 + \pi_{ij} \sum_{k>2} \frac{p_k z_k (2-\gamma p_k^{\delta-1})}{1-2z_k}$$

> 0 for $i \neq j$.

This implies that, the Sen-Yates-Grundy variance estimator of the HT estimator under the scheme is always non-negative.

The foregoing discussions clearly indicate that the scheme retains its πps properties and provides a non-negative value of $v(t_{HT})$ without imposing any restriction on the choice of the parameter δ although the revised probability P_i itself is a function δ. Hence, for different $\delta -$values, the scheme is capable of producing a class of IPPS sampling schemes for selecting two units from the population.

4. SOME SPECIFIC CASES OF S_δ

For some specific selected values of δ, we present corresponding $z_i, \gamma, P_i, P_{j/i}$ and the sampling scheme in Table 1 to show that the IPPS sampling schemes due to Midzuno (1952) and Brewer (1963), and those due to Sahoo *et al.* (2005, 2006) and Senapati *et al.* (2006) developed recently, are particular members of S_δ. But, the domain of S_δ is not restricted only to these noteworthy special cases. Some more such schemes may come out as particular cases of the class for other choices of δ. In the next section, we also provide numerical evaluation of the performance of the scheme for different values of δ.

Table 1: Selected δ – Values and the Resulting Sampling Schemes

Value of δ	z_i	γ	P_i	$P_{j/i}$	Sampling Scheme
0	$\frac{1}{N}$	$\frac{1}{N-1}$	$\frac{2(N-1)p_i-1}{N-2}$	$\frac{1}{N-1}$	Midzuno (1952) (S_M , say)
$+\frac{1}{2}$	$\frac{\sqrt{p_i}}{\sum_{i=1}^{N}\sqrt{p_i}}$	$\frac{\sum_{i=1}^{N}\frac{p_i}{1-2z_i}}{\sum_{i=1}^{N}\frac{\sqrt{p_i}(1-z_i)}{1-2z_i}}$	$\frac{p_i(1-z_i)(2-\gamma/\sqrt{p_i})}{1-2z_i}$	$\frac{z_j}{1-z_i}$	Sahoo *et al.* (2006) (S_1, say)
$+1$	p_i	$\frac{\sum_{i=1}^{N}\frac{p_i}{1-2p_i}}{\sum_{i=1}^{N}\frac{p_i(1-p_i)}{1-2p_i}}$	$\frac{\frac{p_i(1-p_i)}{1-2p_i}}{\sum_{i=1}^{N}\frac{p_i(1-p_i)}{1-2p_i}}$	$\frac{p_j}{1-p_i}$	Brewer (1963) (S_B , say)
-1	$\frac{p_i^{-1}}{\sum_{i=1}^{N}p_i^{-1}}$	$\frac{\sum_{i=1}^{N}\frac{p_i}{1-2z_i}}{\sum_{i=1}^{N}\frac{1-z_i}{p_i(1-2z_i)}}$	$\frac{(1-z_i)(2p_i^2-\gamma)}{p_i(1-2g_i)}$	$\frac{z_j}{1-z_i}$	Senapati *et al.* (2006) (S_2 , say)
$+2$	$\frac{p_i^2}{\sum_{i=1}^{N}p_i^2}$	$\frac{\sum_{i=1}^{N}\frac{p_i}{1-2z_i}}{\sum_{i=1}^{N}\frac{p_i^2(1-z_i)}{1-2z_i}}$	$\frac{p_i(1-z_i)(2-\gamma p_i)}{1-2z_i}$	$\frac{z_j}{1-z_i}$	Sahoo *et al.* (2005) (S_3, say)

5. PERFORMANCE OF S_δ

A desirable further goal is to study efficiency of the proposed sampling scheme for different values of δ compared to some other IPPS sampling procedures. For this purpose, to avoid mathematical difficulties, we have undertaken a numerical study with the help of 7 natural populations as described in Table 2. Fifteen IPPS sampling schemes are taken in to consideration out of which eleven schemes are corresponding to $\delta = 0,\ \pm\frac{1}{2},\ \pm 1,\ \pm 2,\ \pm 3,\ \pm 4$ covering five schemes S_1, S_2, S_3, S_M and S_B defined in Table 1. Four other considered schemes are due to Durbin (1953), Singh (1978), Deshpande and Prabhu Ajgaonkar (1982) and Chao (1982) which we denote by S_4, S_5, S_6 and S_7 respectively. We have not considered IPPS schemes 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 scheme.

Relative efficiency (RE) of the HT estimator under the fifteen competing IPPS sampling schemes, compared to the conventional estimator $\hat{Y}_{pps} = \frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i}$ under PPSWR sampling scheme, are presented in Table 3. Our calculations are based on all $C(N,n)$ possible samples of $n = 2$ drawn from a population.

Table 2: Description of Populations

Pop.	Source	N	y	x
1	Konijn (1973) p.49	16	food expenditure	total expenditure
2	Singh and Chaudhary (1986) p.155	17	no. of milch animals in survey	no. of milch animals in census
3	Yates (1953) p.169	17	area under wheat	total acreage of crops and grass
4	Mukhopadhyay (1998) p.131	12	yield of paddy	area
5	Cochran (1977) p.187	18	population in 1960	population in 1950
6	Jessen (1978) p.151	16	no. of total catch of fish	no. of tagged fish
7	Horvitz and Thompson (1952)	20	no. of households	eye estimated no. of households

An examination of the results shown in Table 3 clearly indicate that the suggested sampling scheme S_δ for all considered values of δ is more efficient than S_4, S_5, S_6 and S_7 in all populations. Although our numerical study is confined to only seven populations, it may lead to a conclusion that the suggested sampling procedure is no way inferior to some standard sampling procedures and can be safely applied in many practical situations.

Table 3: Features of Relative Efficiency of Different IPPS Sampling Schemes

Sampling Scheme		Population						
		1	2	3	4	5	6	7
S_δ	$\delta = -4$	114.30	106.99	109.41	108.82	105.33	111.36	110.92
	$\delta = -3$	111.74	106.90	108.75	109.05	105.71	111.78	111.07
	$\delta = -2$	109.78	106.82	108.54	109.28	105.98	112.12	111.20
	$\delta = -1$ (S_2)	108.41	106.77	108.32	109.48	106.22	112.37	111.31
	$\delta = -\frac{1}{2}$	107.96	106.76	107.98	109.57	106.45	112.44	111.34
	$\delta = 0$ (S_M)	107.58	106.74	107.21	109.62	106.74	112.52	111.37
	$\delta = +\frac{1}{2}$ (S_1)	107.43	106.72	106.98	109.65	106.92	112.57	111.38
	$\delta = +1$ (S_B)	107.28	106.70	106.60	109.68	107.08	112.59	111.40
	$\delta = +2$ (S_3)	107.50	106.75	106.67	109.64	108.65	112.55	111.38
	$\delta = +3$	108.20	106.77	106.80	109.49	108.07	112.41	111.32
	$\delta = +4$	109.36	106.82	107.35	109.24	107.25	112.16	111.21
S_4		101.31	105.84	106.39	106.35	104.49	108.62	109.11
S_5		106.26	106.12	106.41	107.69	104.12	108.31	110.92
S_6		106.25	106.34	106.28	103.95	104.06	109.62	110.84
S_7		106.19	105.90	106.38	108.05	104.34	110.08	109.91

6. SOME REMARKS ON THE OPTIMUM VALUE OF δ

Selection of δ restricts the operation of S_δ because $P_i > 0$ for a finite range of δ only depending on the configurations of $x-$ and $y-$values for the population under consideration. Analytically, it is not possible to trace out an optimum value of δ for which the scheme attains the maximum precision. However, we computed the RE of S_δ compared to the PPSWR scheme for different values of δ using data on a number of populations (artificial and natural) available in text books and research papers on sampling theory. From these computed values as well as those displayed in Table 3, we notice that RE is either a concave or a convex function of δ attaining a minimum or maximum value for a value of $\delta \in [1,2]$. We further computed this performance measure for different values of δ in $[1,2]$. However, we observed that RE is either maximum or minimum for $\delta = 1.1$(approx.).

With the objective of correlating the features of RE for variations in δ with various population characteristics, we also calculated coefficient of variation, skewness and kurtosis of $x-$values of these populations. But we failed to achieve this objective. However, only one thing we noticed that when skewness of x approaches towards zero, RE may be a convex function of δ attaining its maximum value at $\delta = 1.1$ (approx). But, this can not be accepted as a unique criterion for all practical purposes, because our numerical study has a limited scope.

ACKNOWLEDGEMENT
The authors are grateful to the referee for providing some useful comments on an earlier draft of the paper.

RECEIVED OCTOBER, 2009
REVISED JUNE 2010

REFERENCES

[1] 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.

[2] BREWER, K.R.W. and HANIF, M. (1983): **Sampling with Unequal Probabilities**. Lecture Notes in Statistics, Springer-Verlag, Berlin.

[3] CHAO, M. (1982): A general purpose unequal probability sampling plan. **Biometrika,** 69, 653-656.

[4] CHAUDHURI, A. and VOS, J.W.E. (1988): **Unified Theory and Strategies of Survey Sampling**. North Holland, Amsterdam.

[5] COCHRAN, W.G. (1977): **Sampling Techniques**, 3rd Edition. Wiley, New York.

[6] DESHPANDE, M.N. and PRABHU AJGAONKAR, S.G. (1982): An IPPS (inclusion probability proportional to size) sampling scheme. **Statistica Neerlandica**, 36, 209-212.

[7] DURBIN, J. (1953): Some results in sampling theory when the units are selected with unequal probabilities. **Journal of the Royal Statistical Society,** B15, 262-269.

[8] DURBIN, J. (1967): Design of multi-stage surveys for the estimation of sampling errors. **Applied Statistics**, 16, 152-164.

[9] HORVITZ, D.G. and THOMPSON, D.J. (1952): A generalisation of sampling without replacement from a

finite universe. **Journal of the American the Statistical Association**, 47, 663-685.

[10] JESSEN, R.J. (1978): **Statistical Survey Techniques**. Wiley, New York.

[11] KONIJN, H.S. (1973**): Statistical Theory of Sample Survey Design and Analysis**. North Holland, Armestand.

[12] MIDZUNO, H. (1952): On the sampling system with probability proportionate to sum of sizes. **Annals of the Institute of Statistical Mathematics,** 3, 99-107.

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

[14] RAO, J.N.K. (1965): On two simple schemes of unequal probability sampling without replacement. **Journal of the Indian Statistical Association,** 3, 173-180.

[15] SAHOO, L.N., MISHRA, G. and SENAPATI, S.C. (2005): A new sampling scheme with inclusion probability proportional to size. **Journal of Statistical Theory and Applications,** 4, 361-369**.**

[16] SAHOO, L.N., DAS, B.C. and SINGH, G.N. (2006): A note on an IPPS sampling scheme. **Allgemeines Statistisches Archiv,** 90, 385-393.

[17] SAMPFORD, M.R. (1967): On sampling without replacement with unequal probabilities of selection. **Biometrika,** 54, 499-513.

[18] SARNDAL, C.E., SWENSSON, B. and WRETMAN, J. (2003): **Model Assisted Survey Sampling,** 2nd Edition, Springer-Verlag, Berlin.

[19] SEN, A.R. (1953): On the estimator of the variance in sampling with varying probabilities. **Journal of the Indian Society of Agricultural Statistics**, 5, 119-127.

[20] SENAPATI, S.C., SAHOO, L.N. and MISHRA, G. (2006): On a scheme of sampling of two units with inclusion probability proportional to size. **Austrian Journal of Statistics,** 35, 445-454**.**

[21] SINGH, D. and CHAUDHARY, F.S. (1986): **Theory and Analysis of Sample Survey Designs**. Wiley Eastern Limited, New Delhi.

[22] SINGH, P. (1978): The selection of samples of two units with inclusion probabilities proportional to size. **Biometrika,** 65, 450-454.

[23] YATES, F. (1953): **Sampling Methods for Censuses and Surveys**. Charles Griffin & Company, London.

[24] YATES, F. and GRUNDY, P.M. (1953): Selection without replacement from within strata with probability proportional to size. **Journal of the Royal Statistical Society**, B15, 235-261.