Developments in PPS sampling – Impact on current research

by A. R. Sen

Occasional Paper
Number 29

Canadian
Wildlife Service

DEVELOPMENTS IN PPS SAMPLING - IMPACT ON CURRENT RESEARCH

by A.R. Sen

Canadian Wildlife Service
Occasional Paper Number 29

Wildlife Service
Environment Canada

Service de la Faune
Environnement Canada

Issued under the Authority of the
Minister of the Environment

Catalogue No. CW69-1/29
Ottawa

Abstract

This paper reviews the developments in sampling with probability proportional to size (PPS Sampling), discusses their impact on current research and provides references. Recommendations on specific estimators are made for obtaining efficient estimates of population total and its error from the sample values for two situations generally encountered in practice, (a) where strata, not necessarily efficient, are decided in advance e.g., administrative blocks and estimates required for each stratum as well as for the entire population, (b) where estimates are required for the population only and the practitioner can choose how to stratify.

For estimating multiple characters, it is recommended that the classical ratio estimator be used for each character with the auxiliary variate chosen either as that of the character or of a highly correlated one from a previous occasion.

Résumé

La présente publication comporte un survol des progrès en matière d'échantillonnage où le coefficient de probabilité est proportionnel à la taille d'un ensemble (échantillonnage PPS), traite de leur influence sur les recherches en cours et donne des références. On y émet des recommandations au sujet d'estimateurs spécifiques pour l'obtention de nombres estimatifs à écart type minimum de la population totale et du coefficient d'erreur des valeurs échantillonnales dans le cas de deux situations qui surviennent souvent dans la pratique, soit: a) le cas du choix préalable de strates qui ne soient pas nécessairement en fonction d'un écart type minimum, tel celui de l'exigence de subdivisions et nombres estimatifs tant pour chaque strate que pour la population totale; b) le cas où l'on n'exige de nombres estimatifs que de la population, laissant le statisticien disposer de la stratification à sa guise.

Pour l'évaluation simultanée de caractères multiples, on peut employer l'estimateur classique du rapport d'une caractéristique donnée en prenant comme variable auxiliaire ou bien la valeur de cette caractéristique ou bien une qui ait déjà servi et présente un haut coefficient de corrélation.

1. INTRODUCTION

During the last two decades there has been considerable research in sampling with unequal probability and without replacement. A good deal of this work has either been based on results obtained earlier or has been extensions of those results intended to provide more efficient systems for estimating means or totals and their errors. Some of the earlier results have not been referred to in recent publications or texts on sample survey techniques.

The objects of this paper are to review developments in P.P.S. sampling, discuss their impact on current research, and provide references. Single stage selection of clusters (units) from an unstratified population of N clusters will be considered. The methodologies for the stratified and multi-stage cases are straightforward extensions.

2. THEORY

Consider a finite population of N clusters of elements and assume that a sample of n clusters are selected with simple random sampling (SRS) from it. Without loss of generality, we will suppose that two clusters with cluster totals y_1, y_2, clusters means $\bar{y}_1$, $\bar{y}_2$ and cluster sizes x_1, x_2 respectively are selected with SRS out of the population with cluster totals $u_1,\ldots,u_N$ cluster means $\bar{u}_1,\ldots,\bar{u}_N$ and cluster sizes $\nu_1,\ldots,\nu_N$ respectively where $\sum_{i=1}^{N} u_i = Y$, $\sum_{i=1}^{N} \nu_i = X$. According to Neyman (1934) the sample mean

$$\bar{y} = \sum_{i=1}^{2} \frac{y_i}{2} \frac{N}{X} \tag{1}$$

is an unbiased estimate of the population mean of all the elements

$$\bar{Y} = \frac{Y}{X} \tag{2}$$

Presented at the 9th International Biometric Conference held in Boston, Ma. August 22-27, 1976.

The difference between the population quantities u_i, v_i and their sample counterparts y_i and x_i consists in the fact that whereas u_i and v_i are some given specific numbers the values of y_i and x_i vary according to the outcome of the sampling procedure and represent, after the sampling, some of the numbers u_i and v_i according to the clusters that have actually been selected.

Neyman's estimate is usually inefficient since its variance is influenced not only by the variability in cluster means but also by variance in cluster sizes. Since the statistical problem of estimation of total is essentially the same as that of estimation of mean we will henceforth confine the discussion mainly to estimation of total.

2.1 The Ratio Estimator

Workers in U.S.A. developed a number of estimators which though biased had lower mean square error than the Neyman's estimate. A common estimate in vogue is the ratio estimate of the total

$$\hat{Y}_R = \frac{\sum_{i=1}^{2} y_i}{\sum_{i=1}^{2} x_i} X \tag{3}$$

which is consistent.

2.2 Probability Proportional to Size and With Replacement (PPSWR).

If all the x_i's are known, Hansen & Hurwitz (1943) developed a procedure which selects the clusters with probabilities p's proportional to their size (PPS) and with replacement. They also gave a method of selection of the clusters which is based on the cumulative sum of the x_i's. Thus, if there are 3 clusters containing 30, 40 & 20 elements respectively, the first cluster will be assigned as many numbers as are between 1 and 30, the second between

31 and 70 and the third between 71 and 90.

Since sampling with replacement is generally less precise than sampling without replacement unless $\frac{n}{N}$ is small, Hansen & Hurwitz adopted a scheme where the population is divided into a large number of strata and only one cluster is selected from a stratum with PPS and a constant number of elements (sub-units) are sub-sampled with equal probability from the selected clusters. They showed that such systems generally provide marked gains in efficiency in surveys which employ subsampling over systems where the clusters are selected with equal probability. For estimating the variance of the over-all sample mean they advocated grouping the strata in pairs.

For two clusters selected with probability proportional to size and with replacement (PPSWR)

$$\hat{Y}_{PPS} = \frac{1}{2} \sum_{i=1}^{2} \frac{y_i}{p_i} \tag{4}$$

$$v(\hat{Y}_{PPS}) = \frac{1}{2} \sum_{i=1}^{2} \left(\frac{y_i}{p_i} - \hat{Y}_{PPS} \right)^2 \tag{5}$$

where $p_i = x_i/X$, x_i being the size of the i^{th} cluster.

This method suffers from the disadvantages that (1) the same unit may be selected twice, which the practitioner is reluctant to accept, and (2) there is some loss of sampling efficiency. To avoid the first disadvantage Yates & Grundy (1953) suggested that sampling be done without replacement but estimators (4) & (5) be used for estimating population total and variance. The procedure may be objected to on the grounds that both (4) & (5) are biased but working with two examples, Yates and Grundy showed that the bias in the estimated total is trivial. Durbin (1953) showed that when the method of sampling without replacement gives a lower variance than the method of sampling with replacement, the use of the variance formula appropriate to the method of sampling with replacement when selection is made without replacement will lead to an over-estimate of the true variance, the bias being twice (in general, $\frac{n}{n-1}$) the reduction in variance achieved by using PPSWOR sampling instead of PPSWR.

2.3 Unbiased Ratio Estimator.

Hájek (1949), Lahiri (1951), Midzuno (1952) and Sen (1952) developed independently a sampling procedure which amounts to selecting a sample of two clusters with probability proportional to the total of the sizes of the sample clusters ($\sum_{i=1}^{2} x_i$). This makes $\hat{Y}_R$ in (3) an unbiased estimate of Y. Denote this by $\hat{Y}_H$. Hájek, Midzuno[1] and Sen selected one cluster with probability proportional to its size and the other (in general, n-1 clusters) with SRS out of N-1 clusters in the population. Hájek calls this "selection in two phases".

Lahiri's method of selection, however, consists of

(i) selection of two numbers at random, one from 1 to N (say i) and the other from 1 to M (say R), M being the maximum of the sizes of the clusters.

(ii) selection of the i^{th} cluster if $R \leq x_i$,

(iii) rejection of the i^{th} cluster and repetition of the operation if $R > x_i$.

---

1 This method is attributed by Midzuno (1952) to T. Ikeda (1950) who derived it as a special case by putting p = 0 in the selection scheme: (i) select first p clusters with equal probability, (ii) then, select i^{th} cluster out of remaining N-p clusters with probability

$$\frac{\sum_{j=1}^{p} x_j}{(N-p)X} + \frac{x_i}{X}$$

and (iii) finally n-p-1 clusters with SRS.

This procedure leads to the original probabilities of selection of the clusters. But there is a possibility of rejection of certain draws and the probability of rejecting a draw is $1 - \frac{\bar{X}}{M}$ where $\bar{X}$ is the population mean of the sizes.

The probability that a sample with a specified value of Σx_i will be drawn can be shown to be

$$P = \frac{x_i + x_j}{X \binom{N-1}{2-1}} = \frac{x_i \cdot 1}{X.N} + \frac{x_j.}{X} \frac{1}{N} \tag{6}$$

The right hand side of the expression led Hájek, Midzuno & Sen to the selection procedure described above.

Hájek showed that an 'almost unbiased estimate' of $V(\hat{Y}_H)$ is given by

$$v(\hat{Y}_H) = X^2 \left[\frac{s^2(N-2)}{N} + 2(\bar{y}' - \bar{y})^2 \frac{(N-1)}{N} \right] \tag{7}$$

where

$$\bar{y}' = \sum_{i=1}^{2} \bar{y}_i/2; \quad \bar{y} = \sum_{i=1}^{2} y_i / \sum_{i=1}^{2} x_i$$

$$s^2 = \sum_{i=1}^{2} (\bar{y}_i - \bar{y})^2 x_i / \sum_{i=1}^{2} x_i$$

This can be shown

$$= \frac{1}{N}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2 \frac{1}{(p_1 + p_2)^2} \left[(N-2)p_1p_2 + \frac{(N-1)}{2}(p_i - p_2)^2 \right] \tag{8}$$

Raj (1954) and Sen (1952; 1955) have given an unbiased variance estimator

$$v_1(\hat{Y}_H) = \left(\frac{y_1 + y_2}{p_1 + p_2} \right)^2 - \frac{1}{(p_1 + p_2)} \left[(y_1 - y_2)^2 + 2NY_1Y_2 \right] \tag{9}$$

of the variance of the estimate $\hat{Y}_H$ which can take negative values, being generally non-negative for samples with smaller probabilities. Sen (1955) has given a non-negative variance estimator :

$$v_1' = v_1 \text{ if } v_1 \geq 0$$
$$= 0 \text{ if } v_1 < 0 \qquad (10)$$

which is biased. Sen has shown that MSE (v_1') is smaller than that of v_1.

Rao and Vijayan (1976) have given an unbiased estimator of the variance

$$v_2' = \frac{p_1 p_2}{p_1+p_2} \left[(N-1) - \frac{1}{p_1+p_2}\right] \left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 \qquad (11)$$

which is non-negative for samples with larger probabilities and is, therefore, expected to be more efficient than (9).

Neyman's method has the advantage that it is not necessary to know in advance the sizes of the individual clusters and that the average number of selected elements is smaller than it is in the method due to Hájek, Lahiri, Midzuno and Sen.

2.4 Probability Proportional to Size and Without Replacement (PPSWOR).

Horvitz and Thomson (1952) generalized the Hansen & Hurwitz (1943) scheme of selecting 2 or more clusters from a stratum with probability proportional to a measure of size and without replacement (PPSWOR). The HT unbiased estimator of Y (for fixed sample size) for n = 2 is given by

$$\hat{Y}_{HT} = \sum_{i=1}^{2} \frac{y_i}{P_i} \qquad (12)$$

with unbiased variance estimator

$$v_1(\hat{Y}_{HT}) = \sum_{i=1}^{2} \frac{y_i^2(1-P_i)}{P_i^2} + 2y_1y_2 \frac{[P_{12}-P_1P_2]}{P_{12}P_1P_2} \qquad (13)$$

where P_i is the probability of including the i^{th} cluster in a sample of 2 and P_{ij} is the probability that clusters i and j are both in the sample.

It is easy to see

$$P_i = p_i \left[1 + S - \frac{p_i}{1-p_i}\right] \qquad (14)$$

$$\text{where } S = \sum_{i=1}^{N} p_i/1-p_i$$

$$P_{ij} = p_i p_j \left[\frac{1}{1-p_i} + \frac{1}{1-p_j}\right] \qquad (15)$$

As has been stated by Sampford (1975) similar estimates for n = 2 have been considered earlier, for example by Narain (1951).

Another form of the unbiased estimator of variance obtained independently by Sen (1953) and Yates & Grundy (1953) for general n is

$$v_2(\hat{Y}_{HT}) = \frac{(P_1P_2-P_{12})}{P_{12}} \left(\frac{y_1}{P_1} - \frac{y_2}{P_2}\right)^2 \quad \text{for } n = 2 \qquad (16)$$

The estimator (13), though admissible in the class of linear unbiased estimators (Godambe, 1960; Roy & Chakravarti, 1960) in the sense that there does not exist any other member of the class which has a smaller variance than $\hat{Y}_{HT}$, suffers from the weakness that it might take negative values. Sen (1953) showed that $P_{ij} < P_iP_j$ for all i, j for n = 2 and hence (16) is always positive when selection is made with PPS and without replacement using the HT estimator (12) though this is not so for $n \geq 3$ as shown by Singh (1954).

Sen (1953) and Raj (1956) showed that for samples of size n when chosen by the Hájek/Lahiri/Midzuno/Sen method the Sen-Yates-Grundy estimator of variance

$$\sum_{i<j}^{n}\sum \left(\frac{P_iP_j-P_{ij}}{P_{ij}}\right)\left(\frac{y_i}{P_i} - \frac{y_j}{P_j}\right)^2 \qquad (17)$$

is always positive.

For the above scheme, the inclusion probability is given by

$$P_i = \frac{n-1}{N-1} + \frac{N-n}{N-1} \cdot p_i \tag{18}$$

If the y_i are approximately proportional to x_i, it will be far from proportional to P_i (except when $N >> n$), so that the HT estimator will be less precise than the ratio estimator even though it is admissible. In such a case one would prefer the Hájek/Lahiri/Midzuno/Sen form of the estimator instead of $\hat{Y}_{HT}$.

The Sen-Yates-Grundy estimator of variance has been shown to be always positive for any n by Lanke (1974) in rejective sampling (Hájek, 1964) where the sample is selected draw by draw with replacement and the entire sample is accepted if it consists of n distinct units but rejected otherwise when a sample is drawn afresh. For successive sampling (Hájek, 1964) where clusters are drawn with PPSWR until n distinct clusters have been drawn, the corresponding result is true for n = 2 (Lanke, 1974).

Both the estimators (13) & (16) will be subject to large errors (Yates & Grundy, 1953) if some of the P_{ij}'s are extremely small. To deal with this problem Durbin (1967) suggests replacing $\frac{P_iP_j}{P_{ij}} - 1$ by 1 whenever this factor exceeds one. He is of the opinion that the bias in the estimate of variance resulting from this device is negligible in practice. Rao and Singh (1973) show on the basis of empirical evidence that (16) is more stable than (13).

In addition to the consideration that $P_{ij} < P_iP_j$ and $P_{ij} > 0$ for all i, j so that (16) exists and is always positive, an important requirement in PPSWOR is that the HT estimator has to be highly efficient and both (12) & (16) easy to compute. If y_i is exactly proportional to P_i, variance of $\hat{Y}_{HT}$ is zero. Hence, if the values of a 'measure of size' p_i ($\Sigma p_i=1$) are known for all clusters and y_i's are approximately proportional to p_i, $V(\hat{Y}_{HT})$ can be made small by setting P_i proportional to p_i. This is the principle

of I.P.P.S. (inclusion probability proportional to size) schemes, which will not only make P_i easier to compute but will ensure a high efficiency.

The I.P.P.S. methods of Brewer (1963), Fellegi (1963), Carrol and Hartley (1964), Rao (1963,65), Durbin (1967) & Hanurav (1967) use the HT estimator in such a way that cluster i has probability $2p_i$, assumed less than 1, of appearing in the sample. The methods of Brewer (1963) and Rao (1965) are applicable only for n = 2. The methods of Brewer, Durbin and Rao (1963) are equivalent for n = 2 in the sense that their joint probabilities of selection are identical. Sampford's (1967) rejective method is identical with Brewer, Durbin and Rao's (1965) method for n = 2. It is an extension of Durbin's method for n > 2 and is more convenient to use in practice.

In Durbin's method, the first cluster is selected with probability p_i and the second with probability proportional to

$$p_j \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j} \right) \quad j \neq i$$

It is easy to show that

$$P_i = 2p_i \text{ and}$$

$$P_{ij} = p_ip_j \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j} \right) \left(1 + \sum_{k=1}^{N} \frac{p_k}{1-2p_k} \right)^{-1}$$

which are required for estimating total and its variance.

In Sampford's method, the first drawing is made with probability p_i and all subsequent ones with probability proportional to $p_i/(1-np_i)$, all with replacement, the sample being accepted, if it contains n distinct clusters (rejecting completely any sample that does not contain <u>n</u> different clusters).

2.5 Other Developments in PPSWOR.

Raj (1956) considered a set of unbiased ordered estimators in PPSWOR. For samples of size 2, one such estimator

$$\bar{t} = \frac{1}{2} \sum_{i=1}^{2} t_i \qquad (19)$$

where

$$t_1 = \frac{y_1}{p_1} ; \; t_2 = y_1 + \frac{y_2}{p_2} (1-p_1)$$

The estimators t_1 & t_2 are uncorrelated and hence an unbiased variance estimator is given by

$$v(\bar{t}) = \frac{1}{2} \sum_{i=1}^{2} (t_i - \bar{t})^2 \qquad (20)$$

Murthy (1957) has shown that, corresponding to any estimator based on the order of selection of the clusters, there exists a more efficient estimator which ignores the order of selection of the clusters. For two clusters

$$t_u = \frac{1}{(2-p_1-p_2)} \left[(1-p_2)\frac{y_1}{p_1} + (1-p_1)\frac{y_2}{p_2} \right] \qquad (21)$$

$$v(t_u) = \frac{(1-p_1)(1-p_2)(1-p_1-p_2)}{(2-p_1-p_2)^2} \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2 \qquad (22)$$

where $v(t_u)$ is unbiased.

Rao, Hartley and Cochran (1962) have suggested a method of PPSWOR which consists in splitting the population at random into two groups of sizes N_1 and N_2 ($N_1 + N_2 = N$) and a sample of size one is drawn independently from each group with probability $\frac{p_i}{p_i'}$. Their estimator of Y is

$$\hat{Y}_{RHC} = \frac{y_1}{p_1} p_1' + \frac{y_2}{p_2} p_2' \qquad (23)$$

where

$$p_i = \sum_{\text{Group } i} p_i \qquad (i = 1,2)$$

An unbiased estimate of $V(\hat{Y}_{RHC})$ was given by

$$v(\hat{Y}_{RHC}) = C_0 p_1 p_2 \left(\frac{y_1}{p_1} - \frac{y_2}{p_2} \right)^2 \qquad (24)$$

where $C_o = \frac{N-2}{N}$ for N even and $\frac{N-1}{N+1}$ for N odd.

Although the estimator is unbiased, easier to compute and the variance estimator is always positive, it is not generally very efficient since the population is split into groups at random.

3. APPLICATIONS

Rao and Bayless (1969,70) made a useful contribution by comparing (a) the efficiencies of the estimators $\hat{Y}$ of the population total as judged by the inverse of the actual variances and (b) stabilities of the sample estimates of the variances of $\hat{Y}$, as judged by the inverse of the estimates of a group of methods in single-stage sampling. The methods were compared in three situations

(1) 7 small artificial populations,

(2) 20 natural populations to which these methods might be employed and

(3) the much-used super-population model with a linear regression

$$y_i = \beta x_i + e_i, \; i = 1,\dots,N$$

$$E(e_i|x_i) = 0, \quad E(e_i^2|x_i) = ax_i^g$$

$$E(e_i e_j|x_i,x_j) = 0, \quad a > 0 \text{ with } g = 1, 1.5, 1.75, 2$$

The authors presented their results as percent gains in efficiency of the estimators over the Brewer-Durbin methods taken as standard. Their main conclusions are

i) Murthy's method is preferable, when a stable estimator as well as a stable variance estimator is required.

ii) the RHC estimator is the most stable, but it might lead to loss in efficiency.

Cochran (1974) provides an excellent summary of the data analysed by Rao and Bayless. Cochran confined his summary to a slightly smaller group of methods than those used by Rao & Bayless, to natural populations and to the super-population model of Rao & Bayless, using g = 1, 1.5, & 2.

In the natural populations, Cochran found very little difference in average gains in efficiency among the Murthy, RHC and Brewer-Rao and Durbin estimators, the Murthy method proving slightly better than others, the "with replacement" being about 7% poorer. The Brewer-Rao-Durbin method improved as g increased, the rank order at g = 2 being Brewer, Murthy, RHC. Cochran used median values to study percentage gains of the variance estimator over the Brewer-Rao-Durbin variance estimator in view of the highly skewed nature of the distribution. The order of preference in both the natural and the linear models was RHC, Murthy, Brewer-Rao-Durbin.

I have divided the 20 natural populations into two groups (1) populations with $CV(x) \geq 0.7$, $\rho \geq 0.7$ and (2) other populations (see Table 1). To the natural populations, Lahiri & DesRaj methods were added and the DesRaj method was included in the super-population model.

In populations with $\rho \geq 0.7$; $CV(x) \geq 0.7$, Lahiri's method proved highly superior, followed by Murthy, DesRaj & RHC methods; "with replacement sampling" was about 7% inferior. Among 'other populations', Murthy, DesRaj & RHC were about as efficient as Brewer-Rao-Durbin, those due to Lahiri & 'with replacement' proving inferior.

The Brewer-Rao-Durbin method improved with increasing g, the rank order at g = 2 being Brewer, Murthy, DesRaj; the RHC and 'with replacement' proved about 17 percent poorer. In the other group with $CV < 0.7$, $\rho \lesseqgtr 0.7$ (there being no population with $CV \geq 0.7$ and $\rho < 0.7$), Murthy's method was as efficient as Brewer-Rao-Durbin, followed closely by DesRaj & RHC; the 'with replacement' method proved about 8% poorer both for natural as well as super-population model.

The order in percentage gain in efficiency of the variance was RHC, Murthy, DesRaj, 'with replacement' Brewer-Rao-Durbin for natural populations with $\rho \geq 0.7$, $CV(x) \geq 0.7$. For the super-population model the same order was maintained, except that for g = 2 'with replacement' proved 6% inferior to Brewer-Rao-Durbin. In the 'other population group', the order was RHC, Murthy, DesRaj, Brewer; for g = 2 under the linear model, all three proved equally efficient. 'With replacement' proved inefficient both for natural and linear models. There is need for more work on other natural populations to provide firm conclusions.

The Lahiri method proved highly efficient for estimators of natural populations for $\rho \geq 0.7$, $CV(x) \geq 0.7$, though it proved worst for estimating variance. In a personal communication Rao and Vijayan provided an unbiased estimator of the variance of the Lahiri estimator which proved to be highly efficient compared to that by DesRaj and Sen for a number of real populations (N ranging from 8 to 35) and possesses other desirable properties. It would be useful to examine its efficiency relative to Brewer's estimator for the group of natural populations and for the super-population model given in Table 2.

Sampford (1969) has considered samples of 12 from a population of 35 in which y was closely proportional to x over a wide range of x for (i) SRS (unstratified) and (ii) SRS with 4 units drawn from each of 3 strata of sizes 12, 12 and 11, stratified by x. For the SRS unstratified the HT proved the best and the efficiencies relative to the HT estimator were: ratio estimators (biased and unbiased) about 56%, 'with replacement PPS' 54%, Sampford 75%, RHC 77%. For the stratified samples, the advantage was very much reduced and in fact both ratio estimators were more efficient (109%), 'with replacement PPS' 72%, inverse sampling (Sampford, 1961) 89% and RHC 93%.

Hans Stenlund and Anders Westlund (1974) made a Monte Carlo study of 3 sampling designs for 3 different populations (Table 3) with varying degree of skewness (G_1) and excess (G_2). Their main interest was to estimate the target confidence level α based on a confidence interval $\bar{x} \pm Z_{1-\alpha/2} \sqrt{V(\bar{x})}$ where $\bar{x}$ is an estimate of the population mean with the following characteristics (Table 3). Three different sampling procedures were used for each population, (i) SRS, (ii) PPSWOR (DesRaj) and stratified random sampling*, stratification being based on auxiliary variable having 3 different correlations (1, 0.9, 0.6) with the population variable. Stratum bounds were determined according to the principle of cumulative square root frequencies and optimum allocation was used. Stratified sample designs with 2, 3, 4 strata were examined. For the selection procedures SRS & PPSWOR and for each of the 3 populations, 3 different sample sizes 12, 20 and 30 were taken. For each sample size and each population, 500 different samples were drawn.

* This work is due to a student of Stenlund and Westlund.

I have examined the efficiencies of the estimators as judged by the inverse of the actual variances. Stratified random samples with 4 strata provided the best results in most cases, the efficiency being highly marked for P(3). For ρ = 1, 0.9, the design next in order of efficiency was PPSWOR.

For ρ = 0.6, SRS proved better than PPSWOR for P(1) and P(2) but was worst for the highly skewed population P(3) for which stratified random sampling proved highly efficient.

4. CONCLUDING REMARKS

A common problem faced by the survey practitioner today is design of a workable stratified sample design. Two situations generally arise (a) where strata are decided in advance, e.g. administrative blocks, and information is required for each stratum as well as for the population as a whole, (b) where information is required only for the population and the practitioner has the option to employ stratification for increasing efficiency. Assume that we have information available from a previous occasion on the characteristic or a highly correlated characteristic for every cluster (unit) of the population.

Where both estimate of the total & error are required for only one characteristic for situation (a) and one expects a lot of variability between clusters within strata, a choice may be made among Murthy, DesRaj, RHC for each stratum depending upon other features such as ease, flexibility etc. Where it is known before hand that a few clusters are highly variable compared to others, the Hájek-Lahiri-Midzuno-Sen estimator may be used with advantage for estimating the mean (or total) and Rao-Vijayen estimate used for estimating variance.

For situation (b), a stratified sample plan with optimum stratification and allocation is preferred to PPSWOR; its relative efficiency is high in situations where the distribution of the characteristic in the population is likely to be highly skewed. For a normal or near normal population PPSWOR is our best choice.

For estimating a number of characters simultaneously, the classical ratio estimator for a characteristic may be used with the value of the characteristic or of a highly correlated one from a previous occasion used as an auxiliary variable.

I have confined the discussion of PPS sampling to sample size $n = 2$, a situation frequently encountered in practice; extensions are possible to size $n > 2$ and a useful PPSWOR estimator is due to Sampford (1967).

5. REFERENCES

Brewer, K.R.W. (1963). A model for systematic sampling with unequal probabilities - Australian Jour. Statistics, 5, 5-13.

Carrol, J.L. and Hartley, H.O. (1964). The symmetric method of unequal probability sampling without replacement. (Abstract). Biometrics, 20, 908-9.

Cochran, W.G. (1974). Two recent areas of Sample Survey Research, in <u>A Survey of Statistical Design and Linear Models</u>, edited by North Holland.

DesRaj. (1954). Ratio estimation in Sampling with equal and unequal probabilities. Jour. Ind. Soc. Agrl. Stat. 6, 127-138.

DesRaj. (1956). Some estimators in sampling with varying probabilities without replacement. J.A.S.A. 51, 269-284.

Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. Jour. Roy. Stat. Soc., B, 15, 262-269.

Durbin, J. (1967). Estimation of sampling error in multistage surveys. Applied Statistics, 16, 152-164.

Fellegi, I.P. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. J.A.S.A., 58, 183-201.

Godambe, V.P. (1960). An admissible estimator for any sampling design. Sankhya, 22, 285-288.

Hájek, J. (1949). Representativni Vyber Skupin Methodou Dvou Fuzi. Stat. Obzov. 29, (4), 384-394.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population, Ann. Math. Statist. 35, 1491-1523.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. Ann. Math. Stat. 14, 333-362.

Hanurav, T.V. (1967). Optimum utilization of auxiliary information: π PS sampling of two units from a stratum. Jour. Roy. Stat. Soc., Series B, 29, 374-391.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. J.A.S.A. 47, 663-685.

Ikeda, T. (1950). A consideration concerning the selection of samples (in Japanese). Research Memoir of the Institute of Statistical Mathematics 5, 12.

Lahiri, D.B. (1951). A method for sample selection providing unbiased ratio estimates. International Statistical Association Institute Bulletin 33, No. 2, 133-140.

Lanke, Jan (1974). On non-negative variance estimators in survey sampling. Sankhya, Vol. 36, C, 33-42

Midzuno, H. (1952). On the sampling system with probability proportionate to sum of sizes. Ann. Inst. Stat. Math. 2, 99-108.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. Sankhya, 18, 379-390.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Jour. Roy. Stat. Soc., 97, 558-606.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. Jour. Ind. Soc. Agrl. Stat. 3 (1951). 169-175.

Rao, J.N.K., Hartley, H.O., and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. Jour. Roy. Stat. Soc. Series B, 24, 482-491.

Rao, J.N.K. (1963). On three procedures of unequal probability sampling without replacement. J.A.S.A. 58, 202-215.

Rao, J.N.K. (1965). On two simple schemes of unequal probability without replacement. Jour. Ind. Stat. Association 3(1965) 173-180.

Rao, J.N.K. & Bayless, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. J.A.S.A. 64, 540-559.

Rao, J.N.K. and Bayless, D.L. (1970). An empirical study of stabilities of estimators and variance estimators in unequal probability sampling (n = 3 or 4). J.A.S.A. 65, 1645-1667.

Rao, J.N.K. & Singh, M.P. (1973). On the choice of estimator in Survey sampling. Australian J. Statist., 15, 95-104.

Rao, J.N.K. and K. Vijayen. (1976). On estimating the variance in sampling with probability proportional to aggregate size. Unpublished paper, by courtesy of authors.

Roy, J., and Chakravarti, I.M. (1960). Estimating the mean of a finite population. Ann. Math. Statist., 31, 392-398.

Sampford, M.R. (1961). Methods of cluster sampling with and without replacement for clusters of unequal sizes. Biometrika, 49, 27-40.

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

Sampford, M.R. (1969). A comparison of some possible methods of sampling from smallish populations, with units of unequal size - In New Developments in Survey Sampling. Wiley, New York, 170-187.

Sampford, M.R. (1975). The Horvitz-Thompson method in theory and practice - an historical survey. ISI - Warsaw meeting Sept. 1-9.

Sen, A.R. (1952). Present status of probability sampling and its use in estimation of farm characteristics (Abstract). Econometrica, 20, 103.

Sen, A.R. (1952). Further developments of the theory and application of the selection of primary sampling units with special reference to the North Carolina Agricultural Population. Ph.D. thesis Univ. of North Carolina.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. Jour. of the Ind. Soc. Agrl. Statistics 5(1953). 119-127.

Sen, A.R. (1955). On the selection of 'n' primary sampling units from a stratum structure ($n \geq 2$). Annals of Mathematical Statistics, 26, 744-751.

Singh, D. (1954). On the efficiency of the sampling with varying probabilities without replacement. Jour. of the Ind. Soc. Agrl. Statistics 6, 48-57.

Stenlund, H. and Westlund, A. (1974). A Monte Carlo Study of some sampling designs. Statistical Research Report - University of Umeo, Sweden.

Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. Jour. Roy. Stat. Soc. B 15, 253-261.

Table 1. Percent gains in efficiency of the estimators over the Brewer-Rao-Durbin estimator.

	Populations ($\rho \geq 0.7$;CV(x) ≥ 0.7)					Other Populations				
	Murthy	RHC	Lahiri	DesRaj	With Rep.	Murthy	RHC	Lahiri	DesRaj	With Rep.
					Natural Populations					
Mean	6	2	62.5	3.9	-6.5	+0	-1	-5.6	-0.7	-8.0
	[6]	[3]	[16]	[4]	[-5]	[+0]	[-0]	[-3]	[-0]	[-7]
Extremes	(-0,18)	(-1,7)	(-17,511)	(-1,12)	(-17,-1)	(-2,1)	(-6,1)	(-17,4)	(-4,0)	(-17,-5)
					Linear Model g=1					
Mean	5.2	NOT GIVEN		2.2	-6.9	0.6	NOT GIVEN		-0.2	-7.1
Extremes	(1,12)	NOT GIVEN		(0,6)	(-13,-3)	(+0,1)	NOT GIVEN		(-1,0)	(-11,-5)
					Linear Model g=1.5					
Mean	2.1	-5.1	NOT	-0.6	-12.8	0.1	-0.6	NOT	-0.3	-7.8
Extremes	(+0,5)	(-11,-1)	GIVEN	(-2,1)	(-21,-4)	(+0,1)	(-2,-0)	GIVEN	(-1,-0)	(-13,-6)
					Linear Model g=2					
Mean	-1.4	-16.4	NOT	-5.3	-17.3	-0	-1.4	NOT	-0.7	-8.4
Extremes	(-6,-0)	(-23,-2)	GIVEN	(-14,-1)	(-32,-4)	(-0,-0)	(-5,-0)	GIVEN	(-3,-0)	(-15,-6)

1. Figures in [] show corresponding medians.
2. +0 and -0 indicate that the actual values are positive and negative respectively.

Table 2. Percentage gains in efficiency of the variance estimator over the Brewer-Rao-Durbin variance estimator.

	Populations ($\rho \geq 0.7$, CV(x) ≥ 0.7)					Other Populations				
	Murthy	RHC	Lahiri	DesRaj	With rep.	Murthy	RHC	Lahiri	DesRaj	With rep.
					Natural Populations					
Median	22	39	-95	21	8	3	7	-100	3	-6
Quartiles	(8,38)	(15,59)	(-100,-92)	(7,24)	(2,30)	(1,5)	(1,9)	(-100,-86)	(1,4)	(-12,-3)
Extremes	(2,301)	(4,508)	(-100,132)	(2,303)	(-9,322)	(-3,13)	(-5,20)	(-100,-83)	(-3,11)	(-13,0)
					Linear Model g=1					
Median	26	45	NOT	26	20	4	8	NOT	4	-3
Quartiles	(16,54)	(30,85)		(16,52)	(12,54)	(2,8)	(3,16)		(2,8)	(-5,5)
Extremes	(4,277)	(8,433)	GIVEN	(4,268)	(2,284)	(1,17)	(2,31)	GIVEN	(1,17)	(-13,8)
					Linear Model g=1.5					
Median	22	36	NOT	22	14	2	4	NOT	2	-5
Quartiles	(14,68)	(24,97)		(14,65)	(6,58)	(1,4)	(2,6)		(1,4)	(-6,-5)
Extremes	(3,370)	(5,543)	GIVEN	(3,362)	(-1,346)	(1 13)	(1,21)	GIVEN	(1,13)	(-14,-1)
					Linear Model g=2					
Median	13	15	NOT	12	-6	1	1	NOT	0	-9
Quartiles	(7,58)	(8,66)		(5,51)	(-8,29)	(0,1)	(0,1)		(0,1)	(-12,-8)
Extremes	(0,406)	(1,457)	GIVEN	(0,361)	(-9,288)	(0,7)	(0,8)	GIVEN	(0,6)	(-14,-8)

Table 3. Population Characteristics

Population	Size	Mean	Standard Deviation	Skewness G_1	Excess G_2
P(1)	200	2.93	0.97	0.84	0.60
P(2)	200	3.98	3.56	2.17	5.74
P(3)	200	10.47	22.09	4.34	21.76