Chapter 1

Introduction
1.1 Review of Literature

A finite population $U$ of $N$ identifiable units labelled as \{1,...,i,...,N\} is considered. Let $y_i$ denote the value of the characteristic $y$ under study for the $i$th population unit. It is desired to estimate the finite population total $Y = \sum_{i=1}^{N} y_i$ or mean $\overline{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$ on the basis of a random sample. In order to estimate the total or mean of a random sample $s$, the sample units may be selected in two ways:

(1) A pre-determined number $m$ of units may be selected with replacement or

(2) Sampling with replacement may be continued till a desired number $n$ of distinct units is obtained.

The first procedure is called sampling with replacement while the second may be called sampling without replacement.

A comparison is generally made between the two procedures with $m = n$. This comparison, however, is not fair since costs are usually proportional to the number of distinct units and in the first procedure this number would be less than or equal to $n$.

In the first procedure the population mean is usually estimated by the sample mean based on all the units in the sample including repetitions while in the second procedure the estimate is generally made to depend on the distinct units only. Des raj and Khamis (1958) compared the arithmetic mean of the distinct units in the sample $\overline{y}_p = \frac{1}{v} \sum_{i=1}^{v} y_i$, with arithmetic mean of the totality of observed
units, $\bar{y} = \sum_{i=1}^{n} y_i / n$. They showed that $\bar{y}_p$ has smaller variance than $\bar{y}$. Basu (1958) attempted the same comparison and showed that $\bar{y}$ is inadmissible. Basu has used the concept of sufficiency and the Rao-Blackwell theorem on sufficiency to prove $V(\bar{y}_p) \leq V(\bar{y})$. Basu’s argument for the general case of arbitrary probabilities also rests on the idea of sufficiency and he claims that the same results holds. But the concept of sufficiency is not relevant for finite populations. The vector of distinct observations does contain all the information in the sample, but the selection probabilities and possible observational weights necessary for estimation known in advance and independent of the characteristic under study, or determined by counting the appearances of the units, a mere statement of sufficiency does not constitute a proof, unless one is redefining sufficiency.

From the above arguments, should one be restricted to sampling without replacement and forget entirely sampling with replacement? This question has not been answered yet.

If a sample of $n$ units has been drawn from a population of $N$ units, then an unbiased estimator of the population total $Y$ is

$$\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{E(t_i)}$$

in which $E(t_i)$ is the expectation of the number of times that $y_i$ is drawn when taking sample of size $n$.

The first approach in the development of unequal probability sampling has been sampling with replacement and is due to Hansen and Hurwitz (1943). The Hansen and Hurwitz estimator $\hat{Y}_{HH}$ of the population total $Y$ is defined by
\[ \hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^{N} L_i y_i / p_i \]  
\hspace{1cm} (1.1.1) \]

where \( L_i \) denotes the number of times unit \( i \) is selected and \( p_i \) is the probability of selection of unit \( i \) at each of \( n \) draws (\( i = 1, \ldots, N \)).

Basu (1958) presents for this design an unbiased estimator \( \hat{Y}_B = \sum_{i=1}^{u} c_i y_i / p_i \), where the \( c_i \) involve cumbersome computation as \( n \) increases, superior to \( \hat{Y}_{HH} \), which makes use only of the values recorded for the \( u \) distinct units. This estimator \( \hat{Y}_B \) is not identical with

\[ \hat{Y}_d = \frac{1}{u} \sum_{i=1}^{u} \frac{y^{(i)}}{p^{(i)}} \]  
\hspace{1cm} (1.1.2) \]

In fact, \( \hat{Y}_d \) is in general not unbiased. In view of the relative simplicity with which \( \hat{Y}_d \) may be computed, Subrahmanya (1966) studied the properties of \( \hat{Y}_d \) in some detail. The mean-square error of \( \hat{Y}_d \) can be (i) less than that of \( \hat{Y}_d \), (ii) between those \( \hat{Y}_d \) and \( \hat{Y}_{HH} \) or (iii) greater than that of \( \hat{Y}_{HH} \), depending upon the values \( Y_i \) in the population.

The variance of \( \hat{Y}_{HH} \) is

\[ V(\hat{Y}_{HH}) = \frac{1}{n} \sum_{i=1}^{N} p_i \left( \frac{y_i}{p_i} - Y \right)^2 \]

The general theory of probability proportional to size sampling without replacement has been studied by Horvitz and Thompson (1952). The Horvitz–Thompson estimator for the population total is given by
\[ \bar{Y}_{HT} = \sum_{i=1}^{N} L_i \frac{Y_i}{\pi_i} \]

and the variance

\[ V(\bar{Y}_{HT}) = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i} (1 - \pi_i) + \sum_{i \neq j=1}^{N} \frac{Y_i Y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j) \]

where \( L_i \) is as above and \( \pi_{ij} \) denotes the joint inclusion probability of the units \( i \) and \( j \) in the sample and \( \pi_i = \pi_{ii} \). For simple random sampling it is well known that sampling without replacement is more efficient than the with replacement. Rao (1963) has given an example that this need not hold in unequal probability sampling with \( \pi_i = np_i \). One of the disadvantages of unequal probability sampling without replacement, in contrast to sampling with replacement, is the much more cumbersome implementation (See, Särndal, 1996). Therefore, we would at most accept such a procedure if it is combined with a certain gain in efficiency. For fixed sample size Des Raj (1958) prove that

\[ V(\bar{Y}_{HT}) \leq V(\bar{Y}_{HH}) \quad (1.1.3) \]

for all \( y_1, \ldots, y_n \) if, for all \( i \neq j \),

\[ \pi_{ij} \geq (1 - 1/n) \pi_i \pi_j \quad (1.1.4) \]

Unfortunately inequality (1.1.4) is not true for nearly all known procedures without replacement with fixed sample size \( n \). Narain (1951) derived

\[ \pi_{ij} \leq 2(1 - 1/n) \pi_i \pi_j \]
for all \( i \neq j \) as a necessary condition for (1.1.3). For fixed sample size \( n = 2 \) the
Yates-Grundy (1953) variance estimator

\[
\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2, (i \neq j)
\]

will be nonnegative if (1.1.3) holds. \( \pi_{ij} \leq \pi_i \pi_j \), for all \( i \neq j \) does not imply (1.1.3), see Gabler (1984).

Singh (1982) has given the following condition under which sampling without replacement is more efficient than the associated scheme with replacement:

\[
\pi_{ij} \leq (1 - 1/n)(1 - 1/N)^{-1}\pi_i \pi_j \tag{1.1.5}
\]

for all \( i \neq j \). It is easily seen that (1.1.5) can be true only if the sample size \( n \) is fixed and, for all \( i \neq j \),

\[
\pi_{ij} = \frac{n(n-1)}{(N(N-1))}
\]

Thus condition (1.1.5) does not tell us anything new. Some authors proved (1.1.3) for special sampling procedures since condition (1.1.4) of Des Raj is not strong enough. For Sampford’s procedure this was done by Gabler (1981).

Gabler (1984) gives the following condition:

For a connected fixed sample size design with \( \pi_i = np_i > 0 \ (i = 1, \ldots, N) \),

\[
V(\bar{Y}_{HT}) \leq V(\bar{Y}_{HH}) \quad \text{for all} \quad y_1, \ldots, y_N \quad \text{if}
\]

\[
\sum_i \min_j \frac{\pi_{ij}}{\pi_j} \geq n - 1
\]
It should be noted that the Yates-Grundy variance cannot be calculated for sampling designs with \( \pi_{ij} = \pi_i \pi_j \), e.g., Poisson, Promix, Bernoulli sampling etc.

1.2 Sampling with or without replacement

A challenging theoretical problem in unequal probability sampling is the question which design, sampling with or without replacement should be implemented.

(a) Unequal probability sampling with replacement together with estimators based on all the units in the sample has been widely used in practice. This is mainly due to the simplicity of the estimators and their unbiased variance estimators, particularly in multi-stage design when the first stage units (FSUs) are selected with unequal probabilities and with replacement and sub-sampling is done independently from each selected FSU. Also, the past literature on unequal probability sampling without replacement involved certain difficulties such as negative variance estimates and complications when sample size is greater than two. In any case, unequal probability sampling of FSU with replacement together with estimators based on all the FSU in the sample would continue to be useful in multi-stage designs if the survey practitioner desires simple unbiased estimators and variance estimators and is willing to sacrifice some efficiency.

(b) There are business surveys, market surveys for which a pps design is preferable to stratified random sampling from an efficiency point of view (Ålenius, 1990). Furthermore, if the pps sampling procedure is simple to implement, a pps design is simpler to administrate, with much fewer strata to construct, allocate and maintain. Even if we do not claim pps to be generally
preferable to stratified random sampling, it should be considered strong
candidates for business survey designs. For the NASS Crops survey, Bailey
and Kott (1997) found pps sampling to be more efficient than stratified
sampling for most crops.
(c) Särndal (1996) argues for the use of procedures that allow simple, single-sum
variance estimation.
Factors other than efficiency for the choice between the estimators in sampling
with and without replacement are

(1) ** Ease with which a sample can be drawn**

The drawing of a sample by unequal probability sampling without
replacement is not easy to implement except a few sampling designs whereas it is
quite easy to draw a sample with unequal probabilities and with replacement.

(2) **Simplicity of the estimator**

The estimators in sampling with replacement are simple whereas the
estimator with distinct units in sampling without replacement involves
cumbersome computation for sample of size greater than 2 (see, Pathak 1962;
is simple but less efficient than the Horvitz-Thompson (1953) estimator in
sampling without replacement - however, the loss in efficiency will be small if the
sampling fraction is negligible.

(3) **Availability of a variance estimator**

If simple and non-negative unbiased estimates of variance must be had,
clearly the unbiased variance estimator of $\overline{P}_B$ in sampling with replacement is
unsatisfactory since it is complicated and may take negative values. On the other hand, the unbiased variance estimators in sampling without replacement are simple and non-negative. However, in multi-stage designs with several characteristics, even the unbiased estimators in sampling primaries without replacement may prove to be cumbersome since they involve within primary components also. In such situations, the estimator \( \hat{Y} \) (say) based on all the FSUs in sampling FSUs with replacement, though less efficient, may be quite useful since its unbiased variance estimator is given simple by

\[
n(\hat{Y}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} (\hat{Y}_i - \hat{Y})^2
\]

where \( m \) is the number of FSUs in the sample and \( \hat{Y}_i \) is an unbiased estimator of the \( i^{th} \) FSU total \( Y_i \).

(4) Multicharacters

In large-scale sample surveys where the estimators of totals or means of several characteristics are needed, the estimators with distinct units would be highly impractical.

The use of supplementary information is widely discussed issue in sampling theory. Auxiliary variables are commonly used in sample survey practices to obtain improved designs and to achieve higher precision in the estimates of some population parameters such as the mean or the variance of a study variable. This information may be used at the design stage (leading, for instance, to stratification, systematic or probability proportional to size sampling designs), at the estimate stage or both stages. It is well known that when the
auxiliary information is to be used at the estimation stage, the ratio, product and regression methods are widely employed in many situations.

An important milestone with regard to the utilization of multivariate auxiliary information was due to Olkin (1958) who extended the ratio estimator to the case when data on p auxiliary variables is available. This approach has been followed by several authors who developed multivariate product, multivariate difference and multivariate regression estimators. Srivastava (1971) considered a general ratio-type estimator which generates a large class of estimators including most of the estimators up to that time proposed.

Tracy et al. (1996) and Perri (2004), when two auxiliary variables are available, proposed ratio-cum-product estimators. Abu-Dayyeh et al. (2003) introduced two estimators which are indeed members of the class proposed by Srivastava (1971), while Kadilar and Cingi (2004, 2005) analyzed combinations of regression type estimators in the case of two auxiliary variables. In the same situation, Perri (2005) proposed some new estimators obtained from Singh’s (1965, 1967) ones.

1.3 Notations and Terminology

1.3.1 Basic Theory

A survey population U consists of N distinct units identified through the labels \( i = 1, \ldots, N \). The characteristic of interest \( y_i \) associated with unit \( i \) is exactly known by observing the unit \( i \). The value \( y_1, \ldots, y_N \) are fixed, nonrandom real numbers. The random character comes from the randomized sample selection. A sample is a subset, \( s \) of U and the associated \( y \)-values. i.e. \( \{(i, y_i); i \in s\} \) selected
according to a specified sampling design which assigns a known probability sampling design to $s$ such that $p(s) > 0$ for all $s \in \mathcal{S}$ and $\sum_{s \in \mathcal{S}} p(s) = 1$, where $\mathcal{S} = \{s: s \subseteq U\}$, the set of all possible samples.

The first order inclusion probabilities, known for all $i \in U$, are

$$
\pi_i = P(i \in s) = \sum_{s \ni i} p(s)
$$

Here, $s \ni i$ indicates a summation over all the samples $s$ which include unit $i$. We assume that the design is such that $\pi_i > 0$ for all unit $i$.

For computation of variance estimators we need, in addition to the $\pi_i$, to consider second order inclusion probabilities

$$
\pi_{ij} = P(i \& j \in s) = \sum_{s \ni j} \sum_{s \ni i} p(s)
$$

$s \ni (i, j)$ indicates a summation over all the samples $s$ which include both $i$ and $j$. Note that if $i = j$, then $\pi_{ij} = \pi_{ii} = \pi_i$. Let $d_i = \frac{1}{\pi_i}$ and $d_{ij} = \frac{1}{\pi_{ij}}$ denote the design weights of unit $i$ and units $i$ and $j$, respectively. The weights $d_i$ and $d_{ij}$ are fully determined by the sampling design $p(s)$. The designs commonly used are discussed in detail in basic sampling texts, see, e.g. Cochran (1977), Chaudhari and Vos (1988), Särndal et al (1992), Tille (2006), among others.

1.3.2 Estimators and their basic statistical properties

If there is only one study variable $y$, a parameter of interest denoted by $\theta$ is a function of $y_1, \ldots, y_N$, i.e.,
\[ \theta = \theta(y_1, \ldots, y_N) \]

Examples include the population total

\[ \theta = Y = \sum_{i \in U} y_i, \]

the population mean

\[ \theta = \bar{Y} = \sum_{i \in U} \frac{y_i}{N} \]

and the population variance

\[ \theta = S_Y^2 = \sum_{i \in U} \frac{(y_i - \bar{Y})^2}{N - 1} \]

We denote an estimate of \( \theta \) by

\[ \hat{\theta} = \hat{\theta}(S) \]

where \( S \) is a random sample. For example, under SRSWOR, \( \hat{\theta} = \bar{Y} = \sum_{i \in S} y_i \) is often used estimator of the parameter \( Y \).

It is considerable interest to describe the sample-to-sample variations of a proposed estimator \( \hat{\theta} \) for a given design.

The expected value and the variance of \( \hat{\theta} \) are given by

\[ E(\hat{\theta}) = \sum_{S \in \Delta} \hat{\theta}(S)p(S) \]

and

\[ V(\hat{\theta}) = \sum_{S \in \Delta} \{\hat{\theta} - E(\hat{\theta})\}^2 p(S), \]
where \( E_p(\cdot) \) and \( V_p(\cdot) \) denote the design-expectation and design-variance. Thus important measures of the quality of estimation \( \hat{\theta} \) are the bias and the mean square error (MSE). The bias of \( \hat{\theta} \) is defined as

\[
B(\hat{\theta}) = E(\hat{\theta}) - \theta.
\]

An estimator \( \hat{\theta} \) if said to be unbiased for \( \theta \) is \( B(\hat{\theta}) = 0 \) for all \( \underline{y} = (y_1, \ldots, y_N)' \in \mathbb{R}^N \). The MSE of \( \hat{\theta} \) is defined as

\[
MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \sum_{s \in S} \{\hat{\theta} - E(\hat{\theta})\}^2 p(S)
\]

\[
= V(\hat{\theta}) + [B(\hat{\theta})]^2.
\]

**Definition 2.1 (Linear estimator)** An estimator is said to be design linear if it can be written in the form

\[
\hat{\theta} = \sum_{i \in S} w_i y_i
\]

where the \( w_i \) are not functions of the sample \( y_i \)’s.

### 1.4 Various approaches for estimation of finite population parameters

#### 1.4.1 Design-based approach or Probability sampling approach

Probability sampling approach refers to repeated sampling from the survey population \( U \) involving all samples \( s \in S \) and associated probabilities \( p(s) \). It provides valid inferences irrespective of the population \( y \)-values. However, it has been criticized on the ground, the associated inferences, although assumption free refer to repeated sampling instead of just the particular sample \( s \), that has been
drawn. The classical text books by Yates (1949), Deming (1950), Cochran (1953), Hansen, Hurvitz and Madow (1953), Sukhatme (1954) and Särndal (1992), based on the design-based approach, greatly influenced survey practice.

1.4.2 Model-based Approach

The relationship between the auxiliary and survey variables can be used in basically two ways:

1. The traditional design strategy is to use randomized sampling and techniques such as stratification, systematic selection, PPS sampling, ratio and regression estimation. This approach is to state the relationship between the auxiliary and survey variables in terms of the estimator over all possible replications of the population given the super population model. This approach is known as model-design-based approach. Here design-based approach is considered in the light of super population model.

2. It is assumed that the population structure obeys a specified super population model. The distribution induced by the assumed model provides inferences referring to the particular sample of units that has been drawn. Such conditional inferences can be more relevant and appealing than repeated sampling inferences. But model-dependent strategies can perform poorly in large samples when the model is not correctly specified; even small derivations from the assumed model that are not easily detectable through model checking methods can cause serious problems. This approach is also known as Royall’s (1970) approach or Prediction
approach or model-based approach. Here design-related considerations are pushed into the background, design-unbiasedness is no longer considered here.

1.4.3 Model –assisted approach.

A hybrid approach, named model-assisted approach, has been proposed by Särndal, Swensson and Wretman (1989; 1992) and Särndal (1996). In this approach, design-consistent estimators that are also model-unbiased under an assumed “working” model are constructed and used in conjunction with design-consistent variance estimators that are also model-unbiased for the conditional (model) variance. Thus, the resulting inferences are valid under the working model and at the same time protect against model misidentification in the sense of providing valid design-based inferences. The model-assisted approach is certainly useful, but its limitations should also be noted (Chambers, (1996)). The model-assisted approach attempts to combine the desirable features of design-based and model-dependent methods.

In the model-assisted approach, models may be used to construct estimators, but randomization must be used to select the sample, and statistical properties are computed with respect to the probability sampling distribution. There are many examples of the use of model assisted regression estimation in the literature including Estevao, Hindiroglou, and Särndal (1995), Fuller, Loughin, and Baker (1994), and Jayasuriya and Valliant (1996).

Let \((y_i, z_i)\) be pair of values of the study variable \(y\) and an auxiliary variable \(z\) associated with each unit \(i \in U\). Let \(s \subseteq U\) be a sample of size \(n\) drawn
according to probability proportional to size (pps) \( z \), \( p_i \propto z_i \), and with replacement (ppswr) sampling design \( p = p(s) \). Suppose that the values \( y_i, \ i \in s, \) and \( z_i, \ i \in U, \) are known. We seek to estimate the total \( Y = \sum_{i \in U} y_i. \)

### 1.5 Probability Proportional to Size (PPS) Sampling

This is a method of drawing a sample in which the probability \( p_i \) of selecting the \( i \)th population unit is proportional to the size of the unit. The size of the \( i \)th unit is the measure of an auxiliary variable.

**Advantages of using PPSWR sampling:**

**a)** For variable probability sampling schemes without replacement the mean squared error (MSE) of an estimator depends on the joint inclusion probabilities, \( \pi_{ij} \), when \( n > 1 \). The difficulty of designing pps schemes with suitable values of \( \pi_{ij} \)'s is well known and we have not been able to extend the results to any suitable class of ppswr schemes. However, if \( N \) is large and \( n \) is small relative to \( N \) the probability of drawing any unit more than once will be very small and the difference between sampling with and without replacement will be negligible.

**b)** A sampling with probabilities proportional to size is the minimax strategy under fairly strong conditions when the quantity being estimated is not symmetric (see, Scott and Smith, 1975)

**c)** When the study variable and the selection probability, proportional to size measure, are highly correlated, the pps estimator leads to considerable gain in efficiency as compared with the customary estimator under simple random sampling.
(d) The pps estimator is inadmissible. Using Rao-Blackwellization Pathak (1962) obtained a complicated estimator which is less useful in practice than the original pps estimator.

(e) Another advantage of using pps sampling is that no heavy calculations are needed to compute a variance estimator for the pps estimator.

(f) In two-stage sampling one may use ppswr for the first stage and equal probability without replacement for the second stage. If sampling is done without replacement at the first stage, the variance estimator of an estimator of the population mean has two variance components. But if the sampling is done with replacement at the first stage, only one term remains in the variance estimator. In this case, it is simple to estimate variances in multistage sampling with any number of stages when the first stage, with replacement, uses the same unequal probabilities at each drawing while other stages are arbitrary, but carried out independently in different selected first-stage units.

1.6 Research Problem

In survey sampling extra auxiliary information about the finite population is often available at the estimation stage. Utilizing this information more efficient estimators may be obtained. There exist several approaches, such as model-based, calibration, Bayesian etc., each of which provides a practical approach to incorporate auxiliary information at the estimation stage.

Estimation of finite population parameters under PPSWR sampling when extra auxiliary information is available at estimation stage is the main topic of this thesis. The thesis contains seven chapters.
1.7 Chapter wise Summary

Chapter 1. Introduction

In Chapter 1 general introduction about the research problem along with brief literature review is discussed.

Rests of the chapters are as follows.

Chapter 2. PPS Estimation in Presence of Extra Auxiliary Information

In this chapter design-model based approved is used to obtain an optimal estimator of the population total under a general transformation model. The optimal estimator thus obtained depends on the model parameters which are usually not known. Using estimators of the model parameters, selection probabilities are modified and consequently the HH estimator is modified. Further Assuming that extra auxiliary information is available at the estimation stage we have incorporated this information to construct new estimators using ratio and regression method of estimation. A Monte Carlo simulation is carried out to check the performance of the estimators considered in this chapter. From the simulation it has been observed that the suggested estimators have performed well.

Chapter 3. PPS estimation under Multi-characters

In multistage designs, the selection of primaries with probabilities proportional to size, one can use only a single measure of size which is highly positively correlated. It may, sometimes, happen that some of these study
variables are poorly positively correlated with selection probabilities, thereby rendering the existing estimator inadequate.

Rao (1966) proposed some alternative estimators and showed them to be more efficient than the usual estimators. Other references are Rao and Bayless (1969), Bansal and Singh (1985), Kumar and Herzel (1988), Amahice et al. (1989).

In this chapter an attempt was made to suggest alternative estimators by using the information on coefficients of variation of the study and auxiliary variables. A small empirical comparison was presented to study the performance of the suggested estimators.

Chapter 4. A Model-based Estimation of Variance under PPS Sampling

There are two difficulties associated with HT estimator. (i) When both the \( \pi ps \) requirement and the fixed sample size requirement are imposed on the sampling scheme it is tedious and often computationally difficult to calculate \( \pi_{ij} \) satisfying \( \pi_i \pi_j - \pi_{ij} < 0 \). (ii) The design-based variance estimation uses the \( \pi_{ij} \) in a cumbersome double-sum calculation with \( n(n - 1)/2 \) terms. This very large number of terms effectively rules out correct variance calculation to many \( \pi ps \) surveys.

Unequal probability sampling was first suggested by Hansen and Hurwitz (1943) in the context of with-replacement (wr) sampling. Variance estimation for wr sampling requires no heavy calculations. This simplicity speaks in favor of the HH estimator and probability proportional to size (PPS) sampling. The HT estimator with \( \pi ps \)-sampling is identical with the HH estimator with distinct n
units and so, sometimes, one can use HT estimator to estimate the population total or mean and can use variance estimator of the HH estimator to estimate the variance of the HT estimator.

This chapter deals with estimation of the population variance under pps sampling incorporating auxiliary information at estimation stage through model-based approach. An optimal estimator is obtained. Since variance of the HH estimator and population variance are quadratic forms, the estimator of variance of HH estimator could be obtained from the estimator of population variance with the obvious modification (i.e., changing coefficient of square terms and product terms).

Chapter 5. Bayes Estimator and its Admissibility under PPS Sampling.

The pps estimator or Hansen-Hurwitz (HH) estimator depends on multiplicity and hence is inadmissible. This article attempts to construct a generalized difference (GD) estimator and suggests some generalized regression (GREG) estimators under pps sampling. Using the limiting Bayes risk method (see, Lehmann, 1983) admissibility of the GD estimator is established. The relative efficiency of the suggested estimators with respect to the HH estimator has been studied through empirical study.

Chapter 6. Ratio and Regression Methods for PPS Estimator in Two-phase Sampling

In this chapter, using two auxiliary variables $x$ and $z$ we have suggested ratio and regression-type two-phase probability proportional to size (pps)
estimators. The performance of the suggested estimators has been investigated relative to the convention two-phase pps estimator via empirical study is presented.

Chapter 7. Midzuno-Sen Strategy in Two-phase Sampling

This chapter deals with the ratio estimation of the finite population total under Midzuno-Sen sampling scheme when complete auxiliary information is not available but another auxiliary variable, closely related to first auxiliary variable but remotely related to the study variable, is available. A chain ratio estimator in two-phase sampling was suggested. An empirical study was carried out to investigate the performance of the suggested estimator relative to the convention two-phase estimator.

Since the chain ratio estimator under SRSWOR-PPSWR is biased, this sampling procedure is modified and an unbiased ratio-type pps estimator is suggested. Again, since Midzuno-Sen scheme is not IPPS, using transformation of auxiliary variable a modified unbiased Horvitz-Thompson-type estimator is suggested.

1.8 A Monte Carlo Simulation

The precision of an estimator is usually discussed in terms of the variance. When an estimator is a nonlinear function, an exact expression for the variance cannot usually be found. There are some approximate techniques for variance estimation, including random groups, balanced half samples, jackknife, and bootstrap. A more detailed account of these methods is given by Wolter (1985). These techniques are primarily used with more complex estimators. In these
cases, no theoretical results are available, and our knowledge about the statistical properties of the variance estimator is limited to conclusions drawn from simulation studies or other sources of empirical evidence.

In this subsection, we now discuss how one can obtain a simulated bias and variance of a nonlinear estimator.

Monte Carlo refers to a widely used approach for solving complex problems using computer algorithms to simulate the variables in the problem. Typically an algorithm is developed to “model” the problem, and then the algorithm is run many times (from a few hundred up to millions) in order to develop a statistical data set for how the model behaves.

The performance of the various types of estimators will be examined using a Monte Carlo simulation study. We now provide a description of the simulated population and sample creation.

For the empirical comparison of different estimators $\hat{\theta}$ of $\theta$, a sample of size $n$ is drawn using a specific sampling design. The variance estimators are computed from each sample. This process was repeated $M$ times. The performance of the different type of estimators was measured and compared in terms of relative bias in percentage (RB), relative efficiency (RE) and empirical coverage rate (ECR). The simulated values of RB, RE and ECR for a particular variance estimator $\hat{\theta}$ were computed as

$$RB(\hat{\theta}) = 100 \times \frac{\bar{\theta} - \theta}{\theta}$$
where $\bar{\theta}$ is obtained computationally as

$$\bar{\theta} = \frac{1}{M} \sum_{j=1}^{M} \theta_{(j)}$$

The relative percentage standard error and relative efficiency of $\bar{\theta}$ are given by

$$RSE(\bar{\theta}) = \frac{\sqrt{MSE}}{\theta} \times 100$$

and

$$RE(\bar{\theta}) = \frac{MSE(\bar{\theta}_1)}{MSE(\bar{\theta})}$$

where $\bar{\theta}_1$ is the customary estimator of $\theta$ and $MSE(\bar{\theta}) = \frac{1}{M-1} \sum_{j=1}^{M} (\bar{\theta}_j - \theta)^2$.

1.9 Sampling Designs

1.9.1 Simple Random Sampling

This scheme was suggested by Fan, Muller and Rezucha (Fan et al., 1962). This is a least sequential scheme which gives SRSWOR of size $n$.

Let $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ be independent random numbers drawn from the Uniform (0,1) distribution. If $\varepsilon_1 < n/N$, the element $i = 1$ is selected, otherwise not. For subsequent elements, $i = 2, 3, ..., \varepsilon_i$ be the number of elements selected among the first $i - 1$ elements in the population list. If

$$\varepsilon_i = (n - n_i)/(N - i - 1)$$

the element $i$ is selected, otherwise not. The procedure terminates when $n_i = n$. 

23
1.9.2 PPSWR Sampling

Selection of a ppswr sample with each \(i\) is associated a value \(x_i\), the size of unit \(i\). The purpose is to select units with probability proportional to \(x_i\) and \(wr\). This may be done by following ways.

Cumulative total method

A table of cumulative total sizes of the units is made. Let \(T_i = x_1 + x_2 + \cdots + x_i\). A random number, say \(R\), is drawn between 1 and \(T_N (= X)\). The unit \(i\) is selected if \(T_{i-1} < R \leq T_i\). The process is repeated \(n\) times.

1.9.3 Midzuno’s Sampling Scheme (Fixed-size Non-\(\pi\)ps)

The technique of drawing of population elements with unequal probability was first introduced by Hansen and Hurvitz (1943). Later on Horvitz and Thompson (1952) generalized this technique and also considered the problem of estimation. It is noteworthy that the simplest way of incorporating the available supplementary information into a sampling procedure was envisaged by Ikeda, a student of Midzuno (1952) for Midzuno’s estimator as follows:

On the first draw one population element is selected with unequal probabilities and on the second and subsequent draws, elements are selected with equal probabilities without replacement (SRSWOR).

For this selection procedure, the inclusion probabilities for individual and pair-wise units are given by
\[ \pi_t = p_t + \left( \frac{n-1}{N-1} \right) (1-p_t) = \frac{n-1}{N-1} + \frac{N-n}{N-1} p_t \]

and

\[ \pi_{ij} = \frac{n-1}{N-1} \frac{[N-1]}{N-2} \left( \pi_i + \pi_j \right) - \frac{n}{N-2} \]

### 1.10 List of Formulae

\[ \hat{\varphi}_1 = \sum_s \frac{y_i}{n(\bar{z_i})^2} \]

\[ \hat{\varphi}_2 = \sum_s \frac{y_i}{n\{c_x^2(1-p_i) + 1\}(x_i/X) \cdot \sum_U [p_i/\{c_x^2(1-p_i) + 1\}] \}

\[ \hat{\varphi}_3 = \sum_s \frac{y_i}{n\{c_x^2(1-p_i) + 1\}(x_i/X) \cdot \sum_U [p_i/\{c_x^2(1-p_i) + 1\}] \}

\[ \hat{\varphi}_4 = \sum_s \frac{y_i}{n\{c_y c_x/c_x\}^2(1-p_i) + 1\} \frac{x_i}{X} \sum_U [p_i/\{c_y c_x/c_x\}^2(1-p_i) + 1\}] \]

\[ \hat{\varphi}_5 = \sum_s \frac{y_i}{np_{i5}} \]

where, \(p_{i5} = \left\{ \left( c_y + b(c_x - c_x) \right)^2 (1-p_i) + 1 \right\} \cdot \frac{x_i}{X} \cdot \sum_U \left[ p_i/\left( c_y + b(c_x - c_x) \right)^2 (1-p_i) + 1 \right] \]

\[ \hat{\varphi}_{MA1} = \sum_s \frac{y_i}{n(1-p)/N^2} + \rho \{c_x^2(1-p_i) + 1\} \cdot \frac{x_i}{X} \cdot \sum_U [p_i/\{c_x^2(1-p_i) + 1\}] \]

\[ \hat{\varphi}_{MA2} = \sum_s \frac{y_i}{np_{MA2}} \]

where,
\[ p_{M2} = \frac{(1 - \rho)}{N} + \rho \left\{ \left( \frac{c_y c_z}{c_x} \right)^2 (1 - p_i) + 1 \right\} \frac{z_i}{Z}. \]

\[ \sum_U \left[ p_i / \left( \left( \frac{c_y c_z}{c_x} \right)^2 (1 - p_i) + 1 \right) \right] \]

\[ \sigma_{\text{opt}}^2 = \frac{1}{N} \sum_{i \in S} \frac{y_i^2}{N a_i^2} \frac{N a_i^2}{(N - 1) \sum_{l \in U} a_l^2} - \frac{1}{N^2} \sum_{i \neq j} \sum_{c \in S} \frac{y_i y_j}{n(n - 1)} \frac{a_i a_j}{\sum_{l \neq j} \sum_{c \in U} a_l a_j} \]

\[ t^*_L = \sum_{s} \frac{y_i}{np_i} + \left( \sum_{u} a_i - \sum_{s} \frac{a_i}{np_i} \right) \]

\[ \bar{y}_{Rd} = N \left( \frac{\bar{x} \bar{Z}}{\bar{Z}} \right) \sum_s \left( \frac{y_i}{nx_i} \right) \]

\[ \bar{y}_{Regd} = N \left[ \bar{x}' + b' \left( \bar{Z} - \bar{Z}' \right) \right] \sum_s \left( \frac{y_i}{nx_i} \right) \]

\[ \bar{y}_{Dra} = \sum_s \left( \frac{y_i}{np_i} \right) + k \left[ \sum_s \left( \frac{x_i}{np_i} \right) - \sum_s \left( \frac{x_i}{np_i} \right) \right] \]

\[ \bar{y}_{Std} = Z \sum_s \left( \frac{y_i}{nx_i} \right) \sum_{s'} \left( \frac{x_i}{n'x_i} \right) \]

\[ \bar{y}_{rd} = N \left( \frac{\bar{x} \bar{Z}}{\bar{Z}'} \right) \left( \frac{\bar{y}}{\bar{x}} \right) \]

\[ \bar{y}_{adjHT} = \bar{x} \sum_s \frac{y_i}{nx_i} \]

where

\[ \bar{x} = \sum_{s} \frac{w_i}{n_i} - \frac{N(n' - 1)b}{N - n'} \]

and

\[ w_i = x_i + \frac{(n' - 1)b}{N - n'} \]
1.10.1 Estimator Acronyms

FSU - First Stage Units

GD – Generalized Difference

GREG – Generalized Regression Estimators

HH – Hansen-Hurwitz

MS – Midzuno-sen

OPT – Optimal

PPS - Probability Proportional to Size

PPSWR - Probability Proportional to Size With Replacement

SRSWOR – Simple Random Sampling Without Replacement
Paper Presented in Conference /Seminar:

(1) International Conference on Celebrating Statistical Innovation and Impact in a world of Big & Small Data in international Indian Statistical Association (IISA) held during December 19-24, 2015 at Department of Statistics, University of Pune, Pune

(2) National Seminar on Role of Statistics in Current Scenario in Department of Statistics, March 27, 2016 at Department of Statistics, Veer Narmad South Gujarat University, Surat

Training Program Attended:

(1) National Level Short Term Training Program on “Statistical Analysis using R programming” in School of Computer Studies, Ahmadabad University, Ahmadabad, March 07-12, 2016

Research Paper Published:


Research Paper Communicated:

(1) Ratio and regression-type estimation for pps estimator in two-phase sampling, Statistics in Transition (New Series)