Inference from Finite Population Sampling

A Unified Approach

By

Kashmira Ansuyah Hargovan

Inference from Finite Population Sampling

A Unified Approach

By

Kashmira Ansuyah Hargovan

Submitted in fulfilment of the requirements for the degree of Master of Science in the School of Statistics and Actuarial Science at the University of KwaZulu – Natal.

December 2007

Abstract

In this thesis, we have considered the inference aspects of sampling from a finite population. There are significant differences between traditional statistical inference and finite population sampling inference. In the case of finite population sampling, the statistician is free to choose his own sampling design and is not confined to independent and identically distributed observations as is often the case with traditional statistical inference. We look at the correspondence between the sampling design and the sampling scheme. We also look at methods used for drawing samples. The non – existence theorems (Godambe (1955), Hanurav and Basu (1971)) are also discussed. Since the minimum variance unbiased estimator does not exist for infinite populations, a number of estimators need to be considered for estimating the same parameter. We discuss the admissible properties of estimators and the use of sufficient statistics and the Rao-Blackwell Theorem for the improvement of inefficient inadmissible estimators. Sampling strategies using auxiliary information, relating to the population, need to be used as no sampling strategy can provide an efficient estimator of the population parameter in all situations. Finally few well known sampling strategies are studied and compared under a super population model.

Declarations

This dissertation represents original work by the author and has not otherwise been submitted in any form for any degree or diploma to any tertiary institution. Where the use has been made of the work of others, it is duly acknowledged in the text.

_____

K A Hargovan

Acknowledgements

I would like to thank the following people each of whom has made this work possible:

Firstly, I would like to thank my supervisor, Prof R Arnab, for his time and expert guidance.

I would like to extend my gratitude to my co-supervisor, Prof D North for her time and assistance.

A special thanks to Prof J Galpin for her encouragement and support.

My sincerest thanks to my parents, my husband Ved Kavi and to my family for their support, understanding and encouragement.

Finally, I would like to thank my colleagues from both the School of Statistics and Actuarial Science at the University of KwaZulu-Natal and the Department of Statistics and Actuarial Science at Wits University, for their support and assistance.

Contents

Chapter 1

Introduction

Survey sampling is a universally accepted approach for collecting data. Extensive resources are devoted every year for data collation by several government, semi-government and private agencies. There are two generally accepted options for the collection of data. The first option is a study in which every unit of the population is surveyed, called a census. The use of a census to study a population is time consuming, expensive, often impossible and strangely enough, often inaccurate. The other option is to study the characteristics of a population by examining a part of it, this is known as sample survey. The main objective of sample survey is to draw inference on the entire population by surveying a part (sample) of it. The theory of survey sampling has been developed over the past several decades and has provided us with various kinds of reasonable scientific tools for drawing samples and making valid inference about the population parameter of interest. The historical development of survey sampling theory is given by Johnson and Smith (1969), Hansen et al. (1985) and Krishnaiah and Rao (1988) among others.

This thesis presents some inferential aspects when sampling from a finite population only, i.e. when sampling from a finite number of identifiable units.

There are significant differences between inference in the case of finite population sampling and traditional statistical inference, i.e. inference when sampling from the infinite, hypothetical population. In the infinite population setup there is typically a sample of n independent observations $x_1,...,x_n$ on a random variable X with the hypothetical density function $f(x,\theta)$ and the problem is to estimate the unknown parameter θ.

In finite population sampling, the focus is on the actual population of which the sample is a part. In finite population sampling, the statistician is free to choose his own sampling design; that is “man made randomization” is used in selecting a sample. The sampling distribution of a given estimator is therefore something that a statistician creates. Thus in survey sampling, statisticians are not confined to independent and identically distributed observations, as is often the case in traditional statistical inference. The basic concepts, such as parameter, sample, data, estimator, are given a special meaning in survey sampling. Traditional statistical inference and survey sampling inference are not opposing theories, but the special nature of the latter produces some unexpected results. Detailed discussions on this topic are given by Cassel, Särndal and Wretman (1977) and Valliant, Dorfman and Royall (2000) among others.

In this thesis, we discuss some inferential aspects of sampling from finite populations which may be divided into the following categories:
(i) design based inference (ii) model based inference or prediction approach and (iii) model assisted inference

In design based inference, the population vector $\underset{\sim}{y} = (y_1,..,y_N)$ is considered to be fixed. From the population U of size N, a sample s of n units is selected using a sampling design. Here only the y- values belonging to the sample s i.e. $y_i, i \in s$ are observed. The values $y_i, i \notin s$ are thus unknown. We make a link between the observed values $y_i, i \in s$ and unobserved values $y_i, i \notin s$ through the sampling design. The expected behavior of an estimator is the long term average of the performance of an estimator through a hypothetically repeated process of sampling governed by a sampling design chosen. In design based inference some unexpected results may be obtained. The main unexpected result was discussed by Godambe (1955), who proved that in the class of linear homogeneous unbiased estimators, the minimum variance unbiased estimator (MVUE) does not exist. Basu (1971) extended this non-existence result to a wider class of unbiased estimators.

The model based or prediction approach assumes that the population vector $\underline{y}$ is random and obeys a certain model (known as a superpopulation model) and that the model distribution leads to valid inference referring to the particular sample s that has been drawn, irrespective of the sampling design. Once the sample has been drawn, a function of the unobserved random variables generally needs to be predicted. A model of joint probability distributions shows the relationship among the random variables. The probability distribution of the random variables is then used to estimate the desired function of the unobserved random variables. Thus prediction inference is very sensitive to model misspecifications.

The model assisted approach, known as model design based inference is a hybrid of design and model based inference. The advantage of this approach is that it provides valid inferences under a model, enabling valid repeated sampling inferences and at the same time protects against model misspecification.

In both the model based and model-design based approach optimum estimators of the finite population characteristics such as mean, variance etc are available.

In this thesis the relationship between a sampling design and a sampling scheme given in Hanurav's (1966) algorithm is discussed in detail. The details of the non-existence theorems invented by Godambe (1955) and Basu (1971) are also considered extensively. To guard against inefficient estimators, the concept of admissible properties of estimators is discussed. The use of sufficient statistics and the Rao-Blackewell theorem for improving estimators of parameters of a finite population is extensively discussed and optimal sampling strategies under various superpopulation models are investigated. Finally, relative efficiencies of a few well known sampling strategies that are commonly used in practice are studied under a superpopulation model which is frequently used in practice.

Throughout the derivation of the above mentioned results, we assume that the population size N is known, no observational error is present and that the same response rate was achieved.

The thesis is structured as follows:

Chapter 2 consists of definitions and notation which are used throughout the thesis such as population, sampling frame, parameter, sample, sampling design, inclusion probabilities, etc. The consistency conditions of inclusion probabilities and the concept of unbiased estimators for linear and quadratic parametric functions are discussed in this chapter for use in further chapters. The main methods of selection of samples as needed for this study are also discussed viz. the cumulative total method and a sampling design. In this chapter the details of Hanurav's algorithm for the selection of a sample is given along with the theorem regarding the correspondence between sampling design and sampling scheme (Hanurav (1966)). Several examples are provided to show how this algorithm can be used.

In Chapter 3, the concept of various types of linear unbiased estimators is introduced along with suitable examples. The unbiasedness property of the Horvitz-Thompson (1952) estimator is discussed, followed by the derivation of an expression of its variance along with an unbiased estimator of this variance. The concept of a minimum variance unbiased estimator in finite population sampling and the non-existence theorem (Godambe (1955)) are also discussed in this chapter. Modification of Godambe's (1955) results by Hanurav (1966) and the extension of Godambe's results to a wider class of estimators proposed by Basu (1971) make up the remainder of the content of this chapter.

Chapter 4 introduces the concept of admissibility, which may guard against the use of inefficient estimators. Admissibility of the Horvitz-Thompson (1952) estimator in the linear unbiased homogeneous class of estimators is presented. Godambe and Joshi's (1965) results relating to the admissibility of the Horvitz-Thompson (1952) estimator in the class of any unbiased estimator

is then shown. The definition of an inadmissible estimator and the concept of sufficiency in finite population sampling is extensively discussed. The Rao-Blackwellisation technique for improving such inefficient estimators of parameters of a finite population is given with examples. The method of improving:
i) sample mean based on SRSWR sampling,
ii) Hansen-Hurwitz (1943) estimator based on PPSWR sampling and
iii) Raj's (1956) estimator based on PPSWOR sampling
completes the content of this chapter.

In Chapter 5 the concept of the superpopulation model is introduced. Definitions of design unbiased, model unbiased and model-design estimators as well as non-informative sampling design, optimal estimators and optimal strategies are given in this chapter. The model-design or model assisted approach which is a hybrid of the design based and model based approach is also presented. Optimal estimators based on a superpopulation model are then derived. The concept of balancing and robustness as well as optimal design and model unbiased estimators are extensively discussed in this chapter.

Chapter 6: In this chapter we consider some specific sampling strategies and give expressions for the estimation of the population total, its variance and unbiased estimators of the corresponding variance. The relative efficiencies of a few well known sampling strategies that are commonly used in practice are studied under a superpopulation model.

Finally Chapter 7 presents an overall conclusion for this thesis.

Chapter 2

Definitions

In this chapter we have presented some basic notation and definitions such as population, sampling frame, parameter, sample, sampling design etc. that are used throughout this thesis.

We also look at the selection of a sample. When making inference from a population, we select part of the population, known as a sample s, following some suitable sampling design. If $p(s)$, the probability of selection of a particular sample, is equal to one, we call such a sampling design purposive sampling. If $0 < p(s) < 1$, we call such a sampling design probability sampling.

A natural question that arises is how to select a sample given a sampling design when the probabilities of selection of a sample are pre-assigned. There are two popular methods viz. i) the cumulative total method and ii) choosing a sample draw by draw and assigning selection probabilities with each draw. The second method is known as a sampling design. Hanurav (1966) first showed the relationship between the sampling design and sampling scheme. Following Hanurav's (1966) algorithm one can draw a sample which can produce a required sampling design. In this section, we will describe in detail the cumulative total method and Hanurav's algorithm for the selection of a sample.

2.1 Populations

Finite, infinite and continuous populations

A *finite population* is a collection of a finite number of identifiable objects or elements. The elements are called “units” of the population. The total number of elements is known as the size of the population. The population size will be denoted by N.

Examples of finite populations: The number of students in a class, as the number of students is countable and the students are identifiable; similarly the number of houses in a certain locality, etc are examples of finite populations.

Infinite Population

Consider the number of insects in a certain region or the number of bacteria in a test tube, which are very large in number and very difficult to count. These types of populations are referred to as *infinite populations*.
The size of the population N may be known or unknown before a survey. The unknown population size N may sometimes become a subject of interest and may be determined by conducting surveys, such as the estimation of the number of illegal immigrants in a country or estimating the number of animals in a game park.

The population cannot always be identified. For example, if we are selecting a sample of air to measure air pollution, it is not possible to divide the population into identifiable units. Such a population is called a *continuous population*.

In this thesis we will consider finite identifiable populations only. The size of the population N is assumed to be known.

We denote the list of a finite population by

$$U = \{u_1, u_2, ..., u_N\}$$

where u_i, $i \in (1, ..., N)$ is the ith unit of the population and N is the size of the population.

2.2 Sampling Frame

A list of all the units of an identifiable population is called a *sampling frame*.

The sampling frame is the basic material for the selection of a sample. The sampling frame must be complete and up to date i.e. it should not have any omission or duplication of units.

2.3 Parameter

Characteristics of a population are known as *study variables*, these are generally not known before a survey. The study variable will be denoted by y. In a multi-characteristic survey we collect information on more than one variable e.g. In a household survey we might wish to enquire about household income, household expenditure, household size etc. In this case we have several study variables viz. household income, household expenditure and household size.

We let y_i denote the value of a study variable y for the ith unit u_i of a population, then the N-dimensional vector

$$\underline{y} = (y_1,...,y_i,...,y_N)$$

is known as the *parameter* of the population U with respect to the characteristic y.

The *parametric space* is all possible values of the vector $\underline{y}$. Here we consider the parameter space

$$\Omega = (-\infty < y_1 < \infty,...,-\infty < y_i < \infty,...,-\infty < y_N < \infty)$$
$$= R_N$$

where R_N is the N-dimensional Euclidean space (also often referred to as $R \times R \times ... \times R$, N times).

We are generally not interested in knowing the vector $\underline{y}$ but are interested in a function of $\underline{y}$. Such a function of $\underline{y}$ is known as a *parametric function*.

Some commonly used parametric functions of interest are given as follows:

i) $Y = \sum_{i=1}^{N} y_i$, the population total,

ii) $\bar{Y} = \frac{Y}{N}$, the population mean,

iii) $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$, the population variance and

iv) $\frac{S_y}{\bar{Y}}$, the population coefficient of variation.

2.4 Sample

An ordered sequence of elements

$$s = \left(u_{i_1},...,u_{i_j},...,u_{i_{ns}}\right)$$

from a population U is known as a sample where $u_{i_j} \in U$.

All the units of the sample need not be distinct.

The number of units, n_s, including repetition is called the *sample size*.

The total number of distinct units of s is known as the *effective sample size* and is denoted by $v(s)$.

An *ordered sample* shows which draw selects which unit whereas an *unordered sample* contains the distinct units from the ordered sample arranged in ascending order. Thus an unordered sample can be derived from an ordered sample, suppressing the order of selection of the units and their repetition.

Example 2.4

Consider the selection of 4 units from a population of N=5 units

$$U = \left(u_1, u_2, u_3, u_4, u_5\right)$$

where $u_1 \le u_2 \le u_3 \le u_4 \le u_5$.

Let unit u_5 be selected on the first draw, on the second draw unit u_2 is selected, on the third draw unit u_1 is selected and on the fourth draw unit u_2 is selected.

Then the sample $s = (u_5, u_2, u_1, u_2)$ is an ordered sample.

From s we construct an unordered sample $\tilde{s} = (u_1, u_2, u_5)$ by selecting distinct units from s and arranging in ascending order.

2.5 Sampling Design

Let φ be the collection of all possible samples s.
A sampling design p is a function defined on φ satisfying the following conditions:

i) $p(s) \geq 0 \qquad \forall s \in \varphi$

ii) $\sum_{s \in \varphi} p(s) = 1$.

A sampling design is said to be a:

i) Fixed effective size (FES) sampling design if $p\{v(s) = v\}$ = 1
i.e. the number of distinct units $v(s) = v$ is fixed for every sample s with $p(s) > 0$;

ii) Fixed sample size (FSS) sampling design if $p\{n(s) = n\}$ = 1
i.e. the number of units in the sample s is fixed as n.

Example 2.5

Consider a finite population $U = (u_1, u_2, u_3, u_4)$ of N = 4 units.

Let $s_1 = (u_1, u_2, u_2)$, $s_2 = (u_4, u_4)$, $s_3 = (u_1, u_3, u_4)$

and let $p(s_1) = \frac{1}{6},\ p(s_2) = \frac{3}{6},\ p(s_3) = \frac{2}{6}$.

Here $\varphi = (s_1, s_2, s_3)$ and $\sum_{s \in \varphi} p(s) = p(s_1) + p(s_2) + p(s_3) = 1$.

Here (φ, p) forms a sampling design.

$n_{s_1} = 3,\ v_{s_1} = 2;$

$n_{s_2} = 2,\ v_{s_2} = 1;$

$n_{s_3} = 3,\ v_{s_3} = 3.$

2.6 Sampling Scheme

A *sampling scheme* is a method of selection of a sample from a population where units are selected one by one from the population using a pre-assigned set of probabilities of selection of units in each draw.

For a fixed sample size (FSS (n)) design, we assign $p_i(k)$ as the selection probability of the ith unit selected at the kth draw.

The $p_i(k)$'s are subject to

i) $0 \le p_i(k) \le 1 \qquad i = 1,...,N \qquad k = 1,...,n$

ii) $\sum_{i=1}^{N} p_i(k) = 1$ for $k = 1,...,n$.

Remark

Hanurav (1966) stated that any sampling scheme produces a sampling design. There is little difference between the definition of a sampling design

and sampling scheme. The sampling design is a statement of all possible samples and corresponding selection probabilities whereas a sampling scheme is a method of choosing a sample.

2.7 Methods of Selection of Samples

2.7.1 Cumulative Total Method

All possible samples in φ are labelled $s_1,...,s_i,...,s_M$ where M = the total number of samples in φ.

The cumulative total is then calculated:

$$CT_i = p(s_1) + ... + p(s_i) \qquad \text{for } i = 1,...,M .$$

A random number R (say) is then selected, using the Uniform (0,1) Distribution, and a sample s_k is selected if

$$CT_{k-1} < R \leq CT_k \qquad \text{where } CT_0 = 0.$$

Example 2.7

Let $U = \{u_1, u_2, u_3, u_4\}$, we let

$s_1 = (u_1, u_1, u_2)$, $s_2 = (u_1, u_2, u_2)$, $s_3 = (u_3, u_2)$, $s_4 = (u_4)$;

$p(s_1) = 0.25$, $p(s_2) = 0.3$, $p(s_3) = 0.2$ and $p(s_4) = 0.25$.

Table 2.1: Probabilities and cumulative totals for samples s_1 to s_4

s	s_1	s_2	s_3	s_4
$p(s)$	0.25	0.3	0.2	0.25
CT_k	0.25	0.55	0.75	1

Let a random number R = 0.34802 be selected from a uniform population with range (0, 1).The sample s_2 is selected since $CT_1 < R < CT_2$ as CT_1 = 0.25, R = 0.34802 and CT_2 = 0.55.

2.7.2 Hanurav's Algorithm

The most general method of selection of a sample is given by Hanurav (1966) and is known as Hanurav's algorithm.

The algorithm is defined as follows:

$$A = A\{q_1(u_i); q_2(s); q_3(s, u_i)\}$$

where

i) $0 \leq q_1(u_i) \leq 1$, $\sum_{i=1}^{N} q_1(u_i) = 1$ for $i = 1, \ldots, N$

ii) $0 \leq q_2(s) \leq 1$ for $s \in \varphi$, where φ is the possible set of samples that can be defined by this algorithm.

iii) $q_3(s, u_i)$ is defined when $q_2(s, u_i) > 0$ and subject to $0 \leq q_3(s, u_i) \leq 1$,

$$\sum_{i=1}^{N} q_3(s, u_i) = 1 \quad \text{for } i = 1, \ldots, N .$$

Method of selection of a sample:

Step 1:

At the first draw a unit u_{i_1} is selected with probability $q_1(u_{i_1})$.

Step 2:

Here we decide whether the sampling procedure will be terminated or continued. We let $s_{(1)} = u_{i_1}$ be the sample selected in the first draw. A Bernoulli trial with success probability $q_2(s_{(1)})$ is performed. If the trial results in a failure, the sampling procedure is terminated and the selected sample is $s_{(1)} = u_{i_1}$. However if the trial results in a success, we proceed to step 3.

Step 3:

Here we select a second unit u_{i_2} with probability $q_3(s_{(1)}, u_{i_2})$. The selected sample is $s_{(2)} = (u_{i_1}, u_{i_2})$. We then go back to step 2 and perform a Bernoulli trial with success probability $q_2(s_{(2)})$. If the trial results in a failure, the sampling procedure is terminated and the selected sample is $s_{(2)} = (u_{i_1}, u_{i_2})$. However if the trial results in a success, a third unit u_{i_3} is selected with probability $q_3(s_{(2)}, u_{i_3})$ and we let $s_{(3)} = (u_{i_1}, u_{i_2}, u_{i_3})$. The procedure is continued until the process is terminated.

Example 2.8

Let $U = \{1, 2, 3\}$. An example of a sampling algorithm is

$A = A\{q_1(1) = q_1(2) = 0.5, q_2(2) = 0.7, q_2(2,3) = 0.2, q_2(s) = 0$

for the remaining samples in S, $q_3(1/2) = 0.2, q_3(3/2) = 0.8, q_3(1/(2,3)) = 1\}$.

Hanurav (1966) proved one to one correspondence of a sampling design and a sampling scheme as follows:

Theorem

i) Sampling according to Hanurav's algorithm A (in section 2.7.2) results in a sampling design.

ii) For a given sampling design *p*, there exists an algorithm A which results in the design *p*.

Proof:

i) Here we have to show $\sum_{s\in\varphi} p(s) = 1$.

Let S_k be a collection of all samples whose size is k,

then $$\varphi = \bigcup_{k=1}^{n_o} S_k$$

where n_o is the maximum sample size that is required

and $$\sum_{s\in\varphi} p(s) = \sum_{k=1}^{n_o}\sum_{s\in S_k} p(s)\,.$$

Now

$$\sum_{s\in S_1} p(s) = \sum_{i_1=1}^{N} p(u_{i_1}) = \sum_{i_1=1}^{N} q_1(u_{i_1})\,\{1-q_2(u_{i_1})\} = 1 - \sum_{i_1=1}^{N} q_1(u_{i_1})q_2(u_{i_1}) \qquad (2.2.1)$$

$$\sum_{s\in S_2} p(s) = \sum_{i_1=1}^{N}\sum_{i_2=1}^{N} p(u_{i_1},u_{i_2}) = \sum_{i_1=1}^{N}\sum_{i_2=1}^{N} q_1(u_{i_1})\; q_2(u_{i_1},u_{i_2})\,\{1-q_3(u_{i_1},u_{i_2})\}$$

$$= \sum_{i_1=1}^{N} q_1(u_{i_1})q_2(u_{i_1}) - \sum_{i_1=1}^{N}\sum_{i_2=1}^{N} q_1(u_{i_1})\; q_2(u_{i_1},u_{i_2})\; q_3(u_{i_1},u_{i_2}) \qquad (2.2.2)$$

.

.

.

$$\sum_{s\in S_{n_o-1}} p(s)=\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\ldots\sum_{i_{n_o-1}=1} p(u_{i_1},u_{i_2},\ldots,u_{i_{n_o-1}})$$

$$=\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\ldots\sum_{i_{n_o-1}=1} q_1(u_{i_1})\; q_2(u_{i_1})\ldots q_2(u_{i_1},u_{i_2},\ldots,u_{i_{n_o-1}})\{1-q_3(u_{i_1},u_{i_2},\ldots,u_{i_{n_o-1}})\}$$

$$=\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\ldots\sum_{i_{n_o-2}=1} q_1(u_{i_1})\; q_2(u_{i_1})\ldots q_2(u_{i_1},u_{i_2},\ldots,u_{i_{n_o-2}})$$

$$-\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\ldots\sum_{i_{n_o-1}=1} q_1(u_{i_1})\; q_2(u_{i_1})\ldots q_2(u_{i_1},u_{i_2},\ldots,u_{i_{n_o-1}})q_3(u_{i_1},u_{i_2},\ldots,u_{i_{n_o-1}}) \qquad (2.2.3)$$

and

$$\sum_{s\in S_{n_o k}} p(s)=\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\ldots\sum_{i_{n_o k}=1} p(u_{i_1},u_{i_2},\ldots,u_{i_{n_o}})$$

$$=\sum_{i_1=1}^{N}\ldots\sum_{i_{n_o}} q_1(u_{i_1})\; q_2(u_{i_1})q_3(u_{i_1},u_{i_2})q_2(u_{i_1},u_{i_2})\ldots q_2(u_{i_1},u_{i_2},..,u_{i_{n_o}})$$

$$\times\{1-q_3(u_{i_1},u_{i_2},..,u_{i_{n_o}})\} \qquad \text{where } q_3(u_{i_1},u_{i_2},\ldots,u_{i_{n_o}})=0 \qquad (2.2.4)$$

Finally adding $\sum_{s\in S_1} p(s)$, $\sum_{s\in S_2} p(s)$,…, $\sum_{s\in S_{n_o k}} p(s)$ the first part of the theorem is proved.

ii) Here we are given a sampling design p where φ = all possible samples and $p(s)$ is the probability of selection of a sample s $(\in\varphi)$.

We need to show that q_1, q_2 and q_3 can be found so that sampling according to the algorithm $A(q_1,q_2,q_3)$ implements the design.

Let $\varphi_i = \{s / u_{i_1} = u_i\}$ = a collection of samples whose first element is u_{i_1} ;

$\varphi_{ij} = \{s / u_{i_1} = u_i, u_{i_2} = u_j\}$ = collection of samples whose first element is u_i and second element is u_j ; The $\varphi_{j_1..j}$'s are similarly defined.

Let $\beta(i_1, i_2, ..., i_n) = p(u_{i_1}, u_{i_2}, ..., u_{i_n})$

$\alpha(j_1) = \sum_{s \in \varphi_{j_1}} p(s)$

$\alpha(j_1, ..., j_k) = \sum_{s \in \varphi_{j_1 \cdots j_k}} p(s)$ are defined similarly.

Here we check $\varphi = \bigcup_{i=1}^{N} \varphi_i$, $\varphi_i = \bigcup_{j=1}^{N} \varphi_{ij} \bigcup u_i$ etc.

and $\sum_{i=1}^{N} \alpha(i) = 1$, $\sum_j \alpha(i, j) + \beta(i) = \alpha(i)$ etc.

Now following Hanurav (1966), we define:

$q_1(u_{i_1}) = \alpha(i_1)$

$$q_2\left(u_{i_1}, u_{i_2}, ..., u_{i_k}\right) = \begin{cases} 1 - \dfrac{\beta(i_1, i_2, ..., i_k)}{\alpha(i_1, i_2, ..., i_k)} & \text{if } \alpha(i_1, i_2, ..., i_k) \neq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$q_3\left(\left(u_{i_1}, u_{i_2}, ..., u_{i_k}\right), u_{i_{k+1}}\right) = \frac{\alpha(i_1, i_2, ..., i_{k+1})}{\alpha(i_1, i_2, ..., i_{k+1}) - \beta(i_1, i_2, ..., i_{k+1})}$$

if $q_2\left(u_{i_1}, u_{i_2}, ..., u_{i_k}\right) > 0$.

So the probability of drawing a sample $\left(u_{i_1}, u_{i_2}, ..., u_{i_n}\right)$ using the algorithm is

$$p(u_{i_1}, u_{i_2}, ..., u_{i_n}) = q_1(u_{i_1}) \cdot q_2(u_{i_1}) \cdot q_3(u_{i_1}, u_{i_2}) \cdot q_2(u_{i_1}, u_{i_2}) \cdot ... \cdot q_3(u_{i_1}, ..., u_{i_n}) \cdot \left\{1 - q_2(u_{i_1}, ..., u_{i_n})\right\}$$

$$= \alpha(i_1) \cdot \left(1 - \frac{\beta(i_1)}{\alpha(i_1)}\right) \cdot ... \cdot \frac{\alpha(i_1, ..., i_n)}{\alpha(i_1, ..., i_{n-1}) - \beta(i_1, ..., i_{n-1})} \cdot \frac{\beta(i_1, i_2, .., i_n)}{\alpha(i_1, i_2, .., i_n)} .$$

$$= \beta(i_1, i_2, .., i_n)$$

2.7.2.1 Examples using the Algorithm

a) Fixed Sample Size Design

For this sampling scheme, $p\{n_s = n\}$=1. So using the algorithm we get:

$q_1(u_{i_1}) > 0$

$q_2(u_{i_1}) = 1$

$q_3(u_{i_1}, u_{i_2}) \geq 0$

.

. continue this process

.

$q_2(u_{i_1},...,u_{i_{n-1}}) = 1$

$q_3(u_{i_1},...,u_{i_{n-1}}) \geq 0$ continue this until $q_2(u_{i_1},...,u_{i_n}) = 0$.

b) Simple Random Sampling With Replacement (SRSWR)

In this sampling scheme, $p_i(k) = \frac{1}{N}$ which is the probability of selecting the ith unit at the kth draw.

So $q_1(u_{i_1}) = \frac{1}{N} = p_i(1)$

$q_2(u_{i_1}) = 1$

$q_3(u_{i_1}, u_{i_2}) = p_i(1) \cdot p_i(2)$

$= \frac{1}{N} \cdot \frac{1}{N}$

$q_2(u_{i_1}, u_{i_2}) = 1$

$$q_3(u_{i_1}, u_{i_2}, u_{i_3}) = p_i(1) \cdot p_i(2) \cdot p_i(3)$$

$$= \frac{1}{N} \cdot \frac{1}{N} \cdot \frac{1}{N}$$

.

. continue this process

.

$$q_2(u_{i_1}, u_{i_2}, ..., u_{i_{n-1}}) = 1$$

$$q_3(u_{i_1}, u_{i_2}, ..., u_{i_n}) = p_i(1) \cdot p_i(2) \cdot ... \cdot p_i(n)$$

$$= \frac{1}{N} \cdot \ldots \cdot \frac{1}{N}$$

$$= \frac{1}{N^n}$$

$$q_2(u_{i_1}, u_{i_2}, ..., u_{i_n}) = 0$$

So the process stops here.

Example 2.7.2.1

Consider a population of size 20 from which a sample of size 4 is to be selected by the SRSWR method.

Here we associate
Unit 1 with the number 01,
Unit 2 with the number 02,
.
.
Unit 20 with the number 20.

To select a sample of size four, we select a two digit random number from a random number table. If the random number selected is between 01 and 20 inclusive, the corresponding unit is selected. If the random number is greater than 20, no unit is selected.

Using the random number table (Cochran (1977), p19), we get

Random Number	Unit Selected
65	-
18	18
82	-
11	11
10	10
87	-
20	20

So the selected sample is s = {18, 11, 10, 20}.

Where

$$q_1(u_{i_1}) = \frac{1}{N} = \frac{1}{20} = p_i(1)$$

$$q_2(u_{i_1}) = 1$$

$$q_3(u_{i_1}, u_{i_2}) = p_i(1) \cdot p_i(2) = \frac{1}{N} \cdot \frac{1}{N} = \frac{1}{20^2}$$

$$q_2(u_{i_1}, u_{i_2}) = 1$$

$$q_3(u_{i_1}, u_{i_2}, u_{i_3}) = p_i(1) \cdot p_i(2) \cdot p_i(3) = \frac{1}{N} \cdot \frac{1}{N} \cdot \frac{1}{N} = \frac{1}{20^3}$$

$$q_2(u_{i_1}, u_{i_2}, u_{i_2}) = 1$$

$$q_3(u_{i_1}, u_{i_2}, u_{i_3}, u_{i_4}) = p_i(1) \cdot p_i(2) \cdot p_i(3) \cdot p_i(4) = \frac{1}{N} \cdot \frac{1}{N} \cdot \frac{1}{N} \cdot \frac{1}{N} = \frac{1}{20^4}$$

$$q_2(u_{i_1}, u_{i_2}, u_{i_2}, u_{i_4}) = 0 .$$

c) Simple Random Sampling Without Replacement (SRSWOR)

For this sampling scheme,

$p_i(k) = \frac{1}{N-(k-1)}$ if kth unit is not selected in first $k-1$ draws,

$k = 1,...,n$

$= \quad 0$ if kth unit is selected in first $k-1$ draws.

So using the algorithm we get:

$$q_1(u_{i_1}) = \frac{1}{N} = p_i(1)$$

$$q_2(u_{i_1}) = 1$$

$$q_3(u_{i_1}, u_{i_2}) = p_i(1) \cdot p_i(2)$$

$$= \frac{1}{N} \cdot \frac{1}{N-1}$$

$$q_2(u_{i_1}, u_{i_2}) = 1$$

.

.

.

$$q_2(u_{i_1}, u_{i_2}, ..., u_{i_{n-1}}) = 1$$

$$q_3(u_{i_1}, u_{i_2}, ..., u_{i_N}) = p_i(1) \cdot p_i(2) \cdot ... \cdot p_i(n)$$

$$= \frac{1}{N} \cdot \frac{1}{N-1} \cdot \ldots \cdot \frac{1}{N-n+1}$$

$q_2(u_{i_1}, u_{i_2}, ..., u_{i_n}) = 0$ therefore the process stops here.

Example 2.7.2.2

Consider a population of size 20 from which a sample of size 4 is to be selected by the SRSWOR method.

Referring to Example 2.7.2.1, we again use a random number table (Cochran (1977), p19) to select units as follows:

Random Number	Unit Selected
26	-
70	-
15	15
20	20
57	-
76	-
40	-
03	3
20	- not selected as sampling without replacement.
43	-
93	-
48	-
79	-
72	-
12	12

So the selected sample is s = { 15, 20, 3, 12 }.

Where

$$q_1(u_{i_1}) = \frac{1}{N} = \frac{1}{20} = p_i(1)$$

$$q_2(u_{i_1}) = 1$$

$$q_3(u_{i_1}, u_{i_2}) = p_i(1) \cdot p_i(2) = \frac{1}{N} \cdot \frac{1}{N-1} = \frac{1}{20} \cdot \frac{1}{19}$$

$$q_2(u_{i_1}, u_{i_2}) = 1$$

$$q_3(u_{i_1}, u_{i_2}, u_{i_3}) = p_i(1) \cdot p_i(2) \cdot p_i(3) = \frac{1}{N} \cdot \frac{1}{N-1} \cdot \frac{1}{N-2} = \frac{1}{20} \cdot \frac{1}{19} \cdot \frac{1}{18}$$

$$q_2(u_{i_1}, u_{i_2}, u_{i_3}) = 1$$

$$q_3(u_{i_1}, u_{i_2}, u_{i_3}, u_{i_4}) = p_i(1) \cdot p_i(2) \cdot p_i(3) \cdot p_i(4) = \frac{1}{N} \cdot \frac{1}{N-1} \cdot \frac{1}{N-2} \cdot \frac{1}{N-3}$$

$$= \frac{1}{20} \cdot \frac{1}{19} \cdot \frac{1}{18} \cdot \frac{1}{17}$$

$q_2(u_{i_1}, u_{i_2}, u_{i_3}, u_{i_4}) = 0$.

d) Probability Proportional To Size Sampling With Replacement (PPSWR)

For PPSWR sampling the probability of selecting the ith unit at any draw is p_i .

So $q_1(u_{i_1}) = p_{i_1}$

$q_2(u_{i_1}) = 1$

$q_3(u_{i_1}, u_{i_2}) = p_{i_1} \cdot p_{i_2}$

$q_2(u_{i_1}, u_{i_2}) = 1$

$q_3(u_{i_1}, u_{i_2}, u_{i_3}) = p_{i_1} \cdot p_{i_2} \cdot p_{i_3}$

.

.

$q_2(u_{i_1}, u_{i_2}, ..., u_{i_{n-1}}) = 1$

$q_3(u_{i_1}, u_{i_2}, ..., u_{i_N}) = p_{i_1} \cdot p_{i_2} \cdot ... \cdot p_{i_n}$

$q_2(u_{i_1}, u_{i_2}, ..., u_{i_n}) = 0$ so the process stops here.

Example 2.7.2.3

Consider the following data (Cochran (1977), p35), relating to the family income and family size of 10 families:

Table 2.2: Family income and family size of 10 families

Family	1	2	3	4	5	6	7	8	9	10
Income	62	62	87	65	58	92	88	79	83	62
Size	2	3	3	5	4	7	2	4	2	5

We can select a sample of 4 families using PPSWR as follows:
First we need to compute the cumulative totals:

Cum Total	2	5	8	13	17	24	26	30	32	37

We then use a random number table (Cochran (1977), p19) to select units.

Random Number	Unit
40	-
18	6
94	-
44	-
34	10
13	4
11	4

So the selected sample is s = { 6, 10, 4, 4 }.

e) <u>Probability Proportional To Size Sampling Without Replacement (PPSWOR)</u>

In this sampling scheme:

$$p_i(1) = p_{i_1}, \quad p_i(2) = \frac{p_{i_2}}{1-p_{i_1}}, \ \ldots, \ p_i(k) = \frac{p_{i_k}}{1-p_{i_1}-\ldots-p_{i_{k-1}}}.$$

So

$$q_1(u_{i_1}) = p_i(1) = p_{i_1}$$

$$q_2(u_{i_1}) = 1$$

$$q_3(u_{i_1}, u_{i_2}) = p_i(1) \cdot p_i(2) = \begin{cases} p_{i_1} \cdot \dfrac{p_{i_2}}{1 - p_{i_1}} & \text{for } i_1 \neq i_2 \\ 0 & \text{otherwise} \end{cases}$$

$$q_2(u_{i_1}, u_{i_2}) = 1$$

.

.

.

$$q_2(u_{i_1}, u_{i_2}, ..., u_{i_{n-1}}) = 1$$

$$q_3(u_{i_1}, u_{i_2}, ..., u_{i_N}) = p_i(1) \cdot p_i(2) \cdot ... \cdot p_i(n)$$

$$= \begin{cases} p_{i_1} \cdot \dfrac{p_{i_2}}{1 - p_{i_1}} \cdot ... \dfrac{p_{i_n}}{1 - p_{i_1} - ... - p_{i_{n-1}}} & \text{for } i_1 \neq i_2 \neq ... \neq i_n \\ 0 & \text{otherwise} \end{cases}$$

$q_2(u_{i_1}, u_{i_2}, ..., u_{i_n}) = 0$ therefore the process stops here.

Example 2.7.2.4

Referring to example 2.7.2.3 and Table 2.2, we get

Family	1	2	3	4	5	6	7	8	9	10
Cum Total	2	5	8	13	17	24	26	30	32	37

Once again we use random numbers obtained from a random number table (Cochran (1977), p19) to select a sample using PPSWOR.

Random Number	Unit
20	6
05	2
62	-
62	-
96	-
23	-not selected as sampling without replacement.
22	-not selected as sampling without replacement.
48	-
73	-
54	-
73	-
71	-
53	-
32	9
41	-
47	-
60	-
01	1

So the selected sample is s = { 6, 2, 9, 1 }.

f) Midzune-Sen (Midzuno 1952; Sen 1953) Sampling (MS)

The first unit is selected with probability p_i, the remaining $n-1$ units are selected by the SRSWOR method so that:

$$p_i(1) = p_i$$

$$p_i(k) = \frac{1}{(N-1)\ldots(N-k+1)} \qquad \text{for } k = 2,\ldots,n.$$

So $q_1(u_{i_1}) = p_i$

$q_2(u_{i_1}) = 1$

$q_3(u_{i_1}, u_{i_2}) = p_i(1) \cdot p_i(2) = p_{i_1} \cdot \frac{1}{N-1}$

$q_2(u_{i_1}, u_{i_2}) = 1$

.

.

.

$q_2(u_{i_1}, u_{i_2}, ..., u_{i_{n-1}}) = 1$

$q_3(u_{i_1}, u_{i_2}, ..., u_{i_N}) = p_i(1) \cdot p_i(2) \cdot ... \cdot p_i(n)$

$= p_{i_1} \cdot \frac{1}{N-1} \cdot \ldots \cdot \frac{1}{N-n+1}$

$q_2(u_{i_1}, u_{i_2}, ..., u_{i_n}) = 0$ therefore the process stops here.

2.8 Inclusion Probability

The inclusion probability of the unit u_i with respect to the sampling design p is denoted by

$$\pi_i = \sum_{s \supset i} p(s)$$

$$= \sum_{s \in \varphi} I_{si}\, p(s)$$

where $I_{si} = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{if } i \notin s \end{cases}$.

The inclusion probability for the ith and jth units ($i \neq j$) is denoted by

$$\pi_{ij} = \sum I_{si} I_{sj}\, p(s).$$

For simplicity, we write $\pi_{ii} = \pi_i$.

2.9 The consistency conditions of inclusion probabilities (Hanurav (1966))

2.9.1 For any sampling design:

i) $\sum_{i=1}^{N} \pi_i = v = E_p(v(s))$ and

ii) $\sum_{i\neq}^{N}\sum_{j}^{N} \pi_{ij} = Var(v(s)) + v(v-1).$

Proof:

i) $$\sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N}\left(\sum_{s\in\varphi} I_{si}\, p(s)\right)$$

$$= \sum_{s\in\varphi} p(s)\sum_{i=1}^{N} I_{si}$$

$$= \sum_{s\in\varphi} p(s)v(s)$$

$$= v \qquad (2.9.1.1)$$

ii) $$\sum_{i\neq}^{N}\sum_{j}^{N} \pi_{ij} = \sum_{i=1}^{N}\left(\sum_{j(\neq i)}\sum_{s\in\varphi} I_{si} I_{sj}\, p(s)\right)$$

$$= \sum_{s\in\varphi} p(s)\left\{\sum_{i=1}^{N} I_{si}\left(\sum_{j} I_{sj} - I_{si}\right)\right\}$$

$$= \sum_{s\in\varphi} p(s)\left\{\sum_{i=1}^{N} I_{si}\left(v(s) - I_{si}\right)\right\}$$

$$= \sum_{s\in\varphi} p(s)\left\{v(s)\sum_{i=1}^{N} I_{si} - \sum_{i} I_{si}^2\right\}$$

$$= \sum_{s\in\varphi} p(s)\left\{v(s)^2 - v(s)\right\}$$

$$= \sum_{s} v(s)^2\, p(s) - \sum v(s)\, p(s)$$

$$= Var(v(s)) + v(v-1). \qquad (2.9.1.2)$$

2.9.2 For a fixed effective size n sampling design:

iii) $\sum_i \pi_i = n$,

iv) $\sum_{j \neq i} \pi_{ij} = (n-1)\pi_i$ and

v) $\sum\sum_{i \neq j} \pi_{ij} = n(n-1)$.

Proof:

iii) Using result (1.9.1.1) above,

$$E[v(s)] = \sum_s v(s)\,p(s)$$

$$= n\sum_s p(s) = n.$$

So we get

$$\sum_i \pi_i = E[v(s)] = n. \qquad (2.9.2.1)$$

iv) $$\sum_{j \neq i} \pi_{ij} = \sum_{j \neq i}\sum_s p(s) I_{si} I_{sj}$$

$$= \sum_s p(s) I_{si} \sum_{j(\neq 1)} I_{sj}$$

$$= \sum_s p(s) I_{si} \left(n - I_{si}\right)$$

$$= n\sum_s p(s) I_{si} - \sum_s p(s) I_{si}$$

$$= n\pi_i - \pi_i$$

$$= (n-1)\pi_i \qquad (2.9.2.2)$$

v) For a fixed effective size n sampling design, $P\{v_s = n\} = 1$, hence $V(v_s) = 0$.

So using (1.9.1.2) above, we get

$$\sum\sum_{i \neq j} \pi_{ij} = n(n-1). \qquad (2.9.2.3)$$

2.10 Data

The information related to units selected in a sample and its y-value obtained from the survey is known as *data* and is denoted by $d = \{(i, y_i), i \in s\}$.

2.11 Estimator

An *estimator* $T(s, y)$ is a real-valued function $t(d)$, which is free of y_i for $i \notin s$ but may involve y_i for $i \in s$.

The numerical value of an estimator for a given sample is called an *estimate*.

2.11.1 Unbiased Estimators

An estimator $T = T(s, y)$ is said to be a design unbiased estimator or simply unbiased for a population parameter θ if and only if

$$E_p(T) = \sum_{s \in \varphi} T(s, y)\, p(s) = \theta \qquad \forall y \in R_N$$

where E_p denotes the expectation with respect to p and $p(s)$ is the probability of selection of the sample s.

2.11.1.1 Types of unbiased estimators of Y = population total

1. Linear Unbiased Estimator

$$t^* = t^*(s,y) = a_s + \sum_{i \in s} b_{si} y_i$$

$$= a_s + \sum_{i} b_{si} y_i I_{si}$$

where i) $\sum_{i \in s}$ denotes the sum over all distinct units in s

ii) a_s is a constant depending on the sample s and not on the $y_i's$

iii) the $b_{si}'s$ are constants that may depend on the selected sample and the unit i, but is independent of the $y_i's$.

2. Linear Homogeneous Unbiased Estimator

$$\text{t} = t(s,y) = \sum_{i \in s} b_{si} y_i = \sum_{i \in s} b_{si} I_{si} y_i$$

where i) $\sum_{i \in s}$ denotes the sum over all distinct units in s

ii) the $b_{si}'s$ are constants that may depend on the selected sample and the unit i, but is independent of the $y_i's$.

3. Horvitz-Thompson (1952) Estimator

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

The estimator t_{HTE}, called the Horvitz-Thompson Estimator, is defined if π_i is positive for every $i = 1,...,N$.

Note:

Examples of different estimators are given in section 3.1.2.

2.11.1.2 Necessary and sufficient condition for existence of an unbiased estimator:

Theorem 2.11.1 (Hanurav (1966))

A set of necessary and sufficient conditions for the estimability of Y in a given design p is that

$$\pi_i > 0 \qquad \forall i = 1,\ldots,N .$$

Theorem 2.11.2 (Hanurav (1966))

A set of necessary and sufficient conditions for the estimability of the quadratic parametric function

$$Q = l_0 + \sum_{i=1}^{N} l_i y_i + \sum_i q_{ii} y_i^2 + \sum_{i\neq} \sum_j q_{ij} y_i y_j$$

in a design p is given by

i) $\pi_i > 0$ if $l_i^2 + q_{ii}^2 > 0$

or

ii) $\pi_{ij} > 0$ if $q_{ij} + q_{ji} \neq 0$.

Corollary

The variance of an estimator is in general in a quadratic form in $y_i's$, $i = 1,...,N$. For the estimation of the variance, the necessary and sufficient condition of estimability is

$$\pi_{ij} \geq 0 .$$

For a systematic sampling scheme, $\pi_{ij} = 0$ for some $i \neq j$. Here the elements are grouped into clusters and a selection is made where a cluster is chosen to become the sample, the result is that the variance of the sample mean cannot be unbiasedly estimated by using a single systematic sampling design.

Example 2.11

Let N = 9 and n = 3, then the possible systematic samples are:
$s_1 = (1,4,7)$, $s_2 = (2,5,8)$ and $s_3 = (3,6,9)$.

Here $\pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \pi_6 = \pi_7 = \pi_8 = \pi_9 = \frac{1}{3}$

and $\pi_{14} = \pi_{17} = \pi_{47} = \pi_{25} = \pi_{28} = \pi_{58} = \pi_{36} = \pi_{39} = \pi_{69} = \frac{1}{3}$.

The rest of the $\pi_{ij}'s$ viz:

π_{12}, π_{16} etc are equal to zero as element 1 and 2 cannot both be in a sample, similarly neither can element 1 and 6, etc.

Hence from the systematic sampling design, the population mean can be estimated but the variance of the sample mean cannot be estimated because some of the $\pi_{ij}'s$ viz. π_{12}, π_{16} etc are equal to zero.

2.11.1.3 Uniformly Minimum Variance Unbiased Estimator (UMVUE)

T_o, an unbiased estimator of parametric function θ, is called an UMVUE for estimating parametric function θ, if for any other unbiased estimator $\tilde{T}$ ($\neq T_o$), the following conditions are satisfied:

i) $V_p(T_o) \leq V_p(T)$ $\forall y \in R_N$

ii) $V_p(T_o) < V_p(T)$ for at least one $y \in R_N$.

2.11.1.4 Unicluster Sampling Design (Hanurav (1966))

A design $\tilde{p}$ is a unicluster design if any two samples $s, s^* \in \varphi$ with $\tilde{p}(s)$, $\tilde{p}(s^*)$ > 0 imply either

i) $s \cap s^* = \phi$

or

ii) the samples s and s^* are equivalent,

where ϕ is a null set.

2.11.2 Admissible Estimators

An estimator T is said to be admissible in a class C of estimators if there does not exist any other estimator in the class C that is better than T.

i.e. there does not exist an alternate estimator $T^*(\neq T)$ for which the following inequalities hold:

i) $V_p(T^*) \leq V_p(T)$ $\forall T^*(\neq T)$ and $y \in R_N$

ii) $V_p(T^*) < V_p(T)$ for at least one $y \in R_N$.

By using the Rao-Blackwell theorem one can improve an inadmissible estimator using a sufficient statistic. Such a technique is known as Rao-Blackwellization.

2.12 Sampling Strategy

This is a combination of sampling design p and an estimator based on a sample selected using the design p .

2.13 List of Abbreviations used

FES	fixed effective size
FSS	fixed sample size
UMVUE	Uniformly minimum variance unbiased estimator
MVUE	Minimum variance unbiased estimator
SRSWR	Simple random sampling with replacement
SRSWOR	Simple random sampling without replacement
PPSWR	Probability proportional to size sampling with replacement
PPSWOR	Probability proportional to size sampling without replacement
IPPS or πPS	Inclusion probability proportional to size sampling design
BLUP	Best linear unbiased predictor
t_{HTE}	Horvitz-Thompson estimator

t_{RHC} Rao-Hartley-Cochran estimator

t_{MS} Midzuno-Sen estimator

2.14 Conclusion

In this chapter we looked at some definitions and results which will be used in later chapter. Some of these definitions may be repeated in later chapters if they are needed.

We also looked at the selection of samples using the cumulative total method. It should be noted in this method we need to list all possible samples along with their probabilities. In practice it is very difficult to use the cumulative total method if N and n are quite large. If for example N=20 and n=5, we have to list $\binom{20}{5} = 15504$ samples with their probabilities which is very difficult.

Hanurav's algorithm can be used easily and be terminated after a finite number of steps. There are several other popular sampling designs such as the Inclusion Probability Proportional to Size (IPPS or πPS) sampling design and the Rao-Hartley-Cochran sampling design which is available in the literature and which will be discussed in Chapter 6. The natural question to ask is, among all the sampling designs (schemes), which is better or which should ideally be used. The answer to this question will be given in Chapter 3.

Chapter 3

Methods of Estimation

In Chapter 2, we have discussed different methods of sample selection. In this chapter we consider design based inference where the vector $\underline{y} = (y_1,...,y_N)$ is fixed. We assume that if unit i belongs to a sample s, then y_i can be observed without error. In this approach the stochastic element upon which inference can be based, is the one introduced through sampling design. Details are given by Cassel, Särndal and Wretman (1977) and Chaudhuri (1988). In design based inference, expectation is the long term average of the performance of an estimator t through a hypothetically repeated process of sampling.

We present expressions for the Horvitz-Thompson estimator, its variance and an unbiased estimator for its variance.

The concept of unbiasedness and minimum variance unbiased estimators are presented through a design based approach. The celebrated non-existence theorems of Godambe (1955) and Basu (1971) are also discussed in detail. In particular Godambe (1955) showed that the MVUE does not exist in the class of linear homogeneous unbiased estimators. Hanurav (1966) modified Godambe's result by showing that the MVUE does not exist for non-unicluster design. Basu (1971) generalised Godambe's result by proving that the MVUE does not exist in the class of unbiased estimators.

3.1 Definitions

3.1.1 Data

Data is the information collected on one or more characters of interest from selected units in a sample. It is denoted by d.

If a single characteristic y is of interest then y_i is the value of the character obtained for the ith unit.

The data corresponding to an ordered sample $s = \left(u_{i_1},...,u_{i_k},...,u_{i_{1ns}}\right)$ will be denoted by

$$d(s) = \left\{(i_1, y_{i_1}),...,(i_k, y_{i_k}),...,(i_{ns}, y_{i_{ns}})\right\}.$$

3.1.2 Linear Unbiased Estimator

A real valued function of d, $T(s,y) = T(d)$ is called an estimator when it is used as a calculated approximation for a certain parametric function of interest, $\theta(t)$.

3.1.2.1 Linear Homogeneous Estimator

A linear homogeneous estimator is a real valued function

$$t = t(s,y) = \sum_{i \in s} b_{s_i} y_i$$

where $\sum_{i \in s}$ denotes the sum over the distinct units in s and the

b_{s_i}'s are constant and equal to zero for $i \notin s$. The constant b_{s_i} may depend on the selected sample and the unit i, but are independent of the y_i's.

The class of linear homogeneous unbiased estimators will be denoted by C_{lh}.

Examples

- **The sample mean based on unit repetition**

$$\overline{y_{n_s}} = \frac{\sum_{i \in s} y_i}{n_s}$$

is an example of a linear homogeneous estimator,

where $b_{s_i} = \frac{n_i(s)}{n_s}$

for $n_i(s)$ the number of times the ith unit appears in s.

Another example of a linear homogeneous estimator is the

- **sample mean based on distinct units**

$$\overline{y_s} = \frac{\sum_{i \in s} y_i}{n_s},$$

where $b_{s_i} = \frac{1}{n_s}$.

3.1.2.2 Linear Estimator

A linear estimator is defined as

$$t^* = t^*(s, y) = a_s + \sum_{i \in s} b_{s_i} y_i$$

where a_s is a constant depending on the sample s but not on the y_i's.

The class of linear unbiased estimators will be denoted by C_θ.

Examples

- **The Difference Estimator** is an example of a linear unbiased estimator. We let

$$b_{s_i} = \frac{1}{\pi_i} \text{ and } a_s = X - \sum_{i \in s} \frac{x_i}{\pi_i},$$

where π_i = inclusion probability for the ith unit so the difference estimator is defined as

$$t = \sum_{i \in s} \frac{y_i}{\pi_i} - \left(\sum_{i \in s} \frac{x_i}{\pi_i} - X \right).$$

- **The Regression Estimator** is another example of a linear unbiased estimator where if we let

$$b_{s_i} = \frac{1}{\pi_i} \text{ and } a_s = \beta \left(X - \sum_{i \in s} \frac{x_i}{\pi_i} \right)$$

where β is a known constant, then

$$t = \sum_{i \in s} \frac{y_i}{\pi_i} - \beta \left(\sum_{i \in s} \frac{x_i}{\pi_i} - X \right)$$

is the Regression estimator.

3.1.3 Unbiased Estimator

3.1.3.1 Definition

An estimator $T = T(s, y)$ is said to be an unbiased estimator for a population parameter θ if and only if

$$E_p(t) = \sum_{s \in \varphi} T(s, y) p(s) = \theta \qquad \forall y \in R_N$$

where E_p is the expectation with respect to the sampling design p and $p(s)$ is the probability of the selection of a sample s according to design p.

3.1.3.2 Condition of Unbiasedness

A linear homogeneous unbiased estimator

$$t(s, y) = \sum_{i \in s} b_{s_i} y_i$$

will be unbiased for the population total Y if and only if

$$E\left[t(s, y)\right] = \sum_{i=1}^{N} y_i \qquad \forall y \in R_N$$

i.e. $$\sum_{s} t(s, y) p(s) = \sum_{i=1}^{N} y_i$$

i.e. $$\sum_{i} y_i \sum_{s \ni i} b_{s_i} p(s) = \sum_{i=1}^{N} y_i \qquad \forall y \in R_N .$$

Now equating the coefficients of y_i, we find that the necessary and sufficient condition of unbiasedness for $t(s, y)$ is

$$\sum_{s \ni i} b_{s_i} p(s) = 1 \qquad \text{for } i = 1,\ldots,N \qquad (3.1.3.1)$$

i.e. $$\sum_{s} b_{s_i} p(s) I_{s_i} = 1 \qquad \text{for } i = 1,\ldots,N .$$

For a linear non-homogeneous unbiased estimator t^*, the necessary and sufficient condition for unbiasedness of the population total Y is

i) $$\sum_{s} a_s p(s) = 0$$

ii) $$\sum_{s \ni i} b_{s_i} p(s) = 1 \qquad \text{for } i = 1,\ldots,N .$$

Examples

We can construct infinitely many unbiased estimators for a given parametric function. For estimation of the population total Y, we choose a b_{s_i} satisfying condition (3.1.3.1) viz. $\sum_{s \ni i} b_{s_i} p(s) = 1$ in various ways as follows:

i) $b_{s_i} = c_i =$ constant

In this case $\sum_{s \ni i} b_{s_i} p(s) = 1 \Rightarrow c \sum_{s \ni i} p(s) = 1$

So that $c\pi_i = 1$

which leads to $c = 1/\pi_i$

and finally $b_{s_i} = 1/\pi_i$

so the estimator is thus

$$t(s,y)=\sum_{i\in s}\frac{y_i}{\pi_i}. \qquad (3.1.3.2)$$

The above estimator (3.1.3.2) is known as the Horvitz-Thompson (1952) estimator.

ii) Noting that $E\left[b_{s_i}\right]=\sum_{s\ni}b_{s_i}\,p(s)=1$ because $b_{s_i}=0$ for $i\notin s$.

We may choose

$$b_{s_i}=\frac{c_{s_i}}{E\left(c_{s_i}\right)} \qquad c_{s_i}=0 \qquad \text{for } i\notin s$$

as an unbiased estimator of Y.

In particular if we choose $c_{s_i}=\nu_s$, then $b_{s_i}=\frac{\nu_s}{E\left(\nu_s\right)}$ and the corresponding estimator is

$$\sum_{i\in s}b_{s_i}y_i=\sum_{i\in s}\frac{\nu_s}{E\left(\nu_s\right)}=\nu_s\frac{\sum y_i}{E\left(\nu_s\right)}.$$

iii) Let $n_i(s)$ be the number of times the ith unit appears in sample s, for a with replacement sampling scheme.

Then we can find an infinite number of unbiased estimators as follows:

$$t_j\left(s,y\right)=\sum_{i\in s}\frac{\left\{n_i\left(s\right)\right\}}{E\left[\left\{n_i\left(s\right)\right\}j\right]}y_j \qquad \text{for } j=1,2,...$$

In this case, $$b_{s_i}=\frac{\left\{n_i\left(s\right)\right\}}{E\left[\left\{n_i\left(s\right)\right\}j\right]} \qquad \text{for } j=1,2,...$$

3.1.3.3 Horvitz-Thompson (1952) estimator (t_{HTE})

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

$$= \sum_{i} \frac{y_i}{\pi_i} I_{s_i}$$

where $I_{s_i} = \begin{cases} 1 \ if \ i \in s \\ 0 \ if \ i \notin s. \end{cases}$

Clearly t_{HTE} is defined when $\pi_i > 0$ for every $i = 1,...,N$.

The Horvitz-Thompson (1952) estimator is an unbiased estimator of the population total Y, to show this we first need the following theorem:

Theorem 3.1

i) $E(I_{s_i}) = \pi_i$ for $i = 1,...,N$,

ii) $V(I_{s_i}) = \pi_i(1-\pi_i)$ for $i = 1,...,N$ and

iii) $Cov(I_{s_i}, I_{s_j}) = \pi_i \pi_j - \pi_{ij}$ for $i \neq j = 1,...,N$.

Proof:

i) $E(I_{s_i}) = \sum_i I_{s_i} p(s)$

$= \pi_i$ (3.1.3.1)

ii) $V(I_{s_i}) = E(I_{s_i}^2) - E(I_{s_i})^2$

$= E(I_{s_i}) - \pi_i^2 \ = \ \pi_i - \pi_i^2$

$= \pi_i(1-\pi_i)$ (3.1.3.2)

iii) $Cov(I_{s_i}, I_{s_j}) = E(I_{s_i}, I_{s_j}) - E(I_{s_i})E(I_{s_j})$

$$= \sum I_{s_i} I_{s_j} p(s) - \pi_i \pi_j$$

$$= \pi_{ij} - \pi_i \pi_j \qquad (3.1.3.3)$$

Theorem 3.2

i) $E(t_{HTE}) = Y$ and

ii) $V(t_{HTE}) = \sum y_i^2 \left(\frac{1}{\pi_i} - 1 \right) + \sum_{i \neq} \sum_j y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right)$

Proof:

Using Theorem 3.1 above, we find

i) $E(t_{HTE}) = \sum_i \frac{y_i}{\pi_i} E(I_{s_i})$

$$= \sum_i \frac{y_i}{\pi_i} \pi_i \qquad \text{using (3.1.3.1)}$$

$$= \sum_i y_i = Y$$

and

ii) $V(t_{HTE}) = V\left(\sum_i \frac{y_i}{\pi_i} I_{s_i} \right)$

$$= \sum_i \frac{y_i}{\pi_i} V(I_{s_i}) + \sum_{i \neq} \sum_j \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} Cov(I_{s_i}, I_{s_j})$$

using (3.1.3.2) and (3.1.3.3)

$$= \sum_i \frac{y_i^2}{\pi_i} (1 - \pi_i) + \sum_{i \neq} \sum_j \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j).$$

Theorem 3.3

For a fixed effective size n design with number of distinct units in a sample is fixed as $v = n$ i.e. $P\{n_s = v\} = 1$, the variance of the Horvitz-Thomson (1952) estimator is given by

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

Proof:

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

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

$$= \sum_{i\neq}\sum_{j}\frac{y_i^2}{\pi_i}\left(\pi_i\pi_j - \pi_{ij}\right) - \sum_{i\neq}\sum_{j}\frac{y_iy_j}{\pi_i\pi_j}\left(\pi_i\pi_j - \pi_{ij}\right)$$

$$= \sum\frac{y_i^2}{\pi_i}\left(\sum_{j\neq i}\pi_j - \sum_{j\neq i}\pi_{ij}\right) - \sum_{i\neq}\sum_{j}\frac{y_iy_j}{\pi_i\pi_j}\left(\pi_i\pi_j - \pi_{ij}\right)$$

Now for a fixed effective size sampling design $P\{n_s = v\} = 1$, we have shown (Chapter 2 equations (2.9.2.1) and (2.9.2.2)) that

$$\sum\pi_i = n \text{ and } \sum_{j\neq i}\pi_{ij} = (n-1)\pi_i ,$$

so we have

i) $$\sum_{j\neq i}\pi_j = \sum_{j}\pi_j - \pi_i$$

$$= n - \pi_i \qquad \text{and}$$

ii) $\sum_{j \neq i} \pi_{ij} = (n-1)\pi_i$.

We thus get that

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

$$= \sum_i \frac{y_i^2}{\pi_i}(1-\pi_i) + \sum_{i\neq}\sum_j \frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{ij} - \pi_i\pi_j)$$

$$= V(t_{HTE})\ .$$

3.1.3.3.1 Inclusion Probability Proportional to size Sampling Design ($IPPS$ or πPS)

Let us put $y_i = c\pi_i$ (where c_i is a constant) in the Horvitz Thompson (1952) estimator expression

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

$$t_{HTE} = c\sum_{i\in s}$$

Now if $\sum_{i\in s}$ = number of distinct units $= v$ which is a constant, then

$t_{HTE} = cv$ which is also a constant.

So the variance of t_{HTE} becomes zero for fixed effective sample size design when $\pi_i \ \alpha \ y_i$.

So if we choose a sampling design for which $\pi_i \ \alpha \ y_i$ then the Horvitz-Thompson (1952) estimator becomes the most efficient in the sense of having the smallest possible variance of zero.

In practice the $y_i's$ are unknown, so we cannot choose π_i's proportional to $y_i's$. However in some situations we may find an auxiliary variable x which is approximately proportional to y. In such a situation we choose π_i to be proportional to x_i.

A sampling design for which inclusion probabilities are proportional to the measure of size (auxiliary) is known as $IPPS$ or πPS sampling design. Obviously πPS sampling design can be implemented if all x_i's are known and positive. Several πPS sampling designs are available in literature. Some of these are discussed in Chapter 6.

3.1.3.4 Minimum Variance Unbiased Estimator (MVUE)

Definitions:

- **Better Estimator**

  Let T_1 and $T_2 (\neq T_1)$ be two unbiased estimators belonging to the class C_θ.

  The estimator T_1 is said to be better than T_2 if

  i) $V_p(T_1) \leq V_p(T_2)$ $\forall y \in R_N$

  and

  ii) The inequality

  $V_p(T_1) < V_p(T_2)$ holds for some $y \in R_N$.

- **MVUE**

An estimator T_o which belongs to C_θ, the class of linear unbiased estimators of θ is called an MVUE for estimating the parametric function θ if T_o is better than any other unbiased estimators belonging to the class C_θ.

i.e. any $\tilde{T}(\neq T_o) \in C_\theta$ satisfies

i) $V_p(T_0) \leq V_p(\tilde{T})$ $\forall y \in R_N$

and

ii) $V_p(T_0) < V_p(\tilde{T})$ for at least one $y \in R_N$.

3.1.3.4.1 Non Existence of MVUE

i) Godambe (1955)

In the class of linear homogeneous unbiased estimators C_{lh}, the MVUE (minimum variance unbiased estimator) does not exist.

Proof:

Let $t(s,y) = \sum_{i \in s} b_{si} y_i$ be a homogeneous linear unbiased estimator for Y.

Then the constants b_{s_i}'s satisfy the unbiasedness condition

$$\sum_{s \supset i} b_{s_i} p(s) = 1 \qquad \text{for every } i = 1,...,N. \qquad (3.1.3.1)$$

Here the objective is to find the constants, b_{s_i} 's that minimize

$$V_p(t) = \sum_{s\in\varphi}(\sum_{i\in s} b_{s_i} y_i)^2 p(s) - Y^2 \qquad (3.1.3.2)$$

subject to the unbiasedness condition (3.1.3.1).

For minimization we consider

$$\psi = \sum_{s\in\varphi}(\sum_{i\in s} b_{s_i} y_i)^2 p(s) - Y^2 - \sum_{i=1}^{N} \lambda_i (\sum_{s\supset i} b_{s_i} p(s) - 1) \qquad (3.1.3.3)$$

where λ_i 's are the undetermined Lagrange multipliers.

Differentiating ψ with respect b_{s_i} and equating to zero, we get

$$\frac{\partial \Psi}{\partial b_{si}} = 2y_i(\sum_{i\in s} b_{s_i} y_i) p(s) - \lambda_i p(s) = 0 \qquad (3.1.3.4)$$

this is equivalent to

$$t(s) = (\sum_{i\in s} b_{s_i} y_i) = \frac{\lambda_i}{2y_i} \qquad \forall\ i \in s,\ y_i \neq 0. \qquad (3.1.3.5)$$

The equation above says that if a sample s contains units i and j ($i \neq j$) we must have

$$t(s,y) = (\sum_{i\in s} b_{s_i} y_i) = \frac{\lambda_i}{2y_i} = \frac{\lambda_j}{2y_j} \qquad \text{for } y_i, y_j \neq 0. \qquad (3.1.3.6)$$

The equation (3.1.3.6) implies that the estimator $t(s,y)$ is independent of the $y_i's$ for $i \in s$. This is impossible. Hence there does not exist a MVUE.

ii) Hanurav (1966)

Hanurav pointed out that Godambe's (1955) result does not relate to unicluster sampling.

Definition: Unicluster Sampling Design

Hanurav defined a design $\tilde{p}$ as a *unicluster design* if any two samples $s, s^* \in \varphi$ with $\tilde{p}(s), \tilde{p}(s^*) > 0$ imply either

i) $s \cap s^* = \phi$ or

ii) $s \sim s^*$.

i.e. either s and s^* are disjoint or they contain the same set of units.

Hanurav (1966) modified Godambe's result as follows:

For a non-census sampling design p with $\pi_i > 0$ for all $i = 1,...,N$, a MVUE does not and does exist in the class C_{lh} of linear unbiased estimators of the population total Y, if p is a non-unicluster and unicluster design respectively.

Proof:

Let p be a non-unicluster design.

Then we must have two samples s_1 and s_2 with $p(s_1)$, $p(s_2)$ > 0 and such that s_1 contains units i and j but not k $(i \neq j \neq k)$ and s_2 contains i and k but not j. In this case $t(s_1, y) = t(s_2, y)t(s_1, y)$ for all non zero values of y_i, y_j and y_k which is impossible because the magnitude of $t(s_1)$ depends on y_i and y_j but

is independent of y_k while $t(s_2)$ depends on y_i and y_k and is independent of y_j. Hence we cannot find constants, b_{si}`s which minimize $V_p(t)$ and satisfy the unbiasedness condition (3.1.3.1).

Now suppose that p is a non-census unicluster design, then $p(s_1)$, $p(s_2)$ > 0 implies that $s_1 \cap s_2 = \phi$ but not $s_1 \sim s_2 \; \forall s_1, s_2 \in \varphi$ because $s_1 \sim s_2 \; \forall s_1, s_2 \in \varphi$ and $\pi_i > 0$ imply that the design p is a census one. So, for a unicluster design p, all the samples must be disjoint and hence a unit can occur only in one sample. Hence the unbiasedness condition $\sum_{s \supset i} b_{s_i} p(s) = 1$ implies $b_{s_i} p(s) = 1$ for every $i \in s$, $i = 1,...,N$.

We thus conclude that for a unicluster, only one unbiased estimator exists,

viz. $$t(s, y) = \frac{\sum_{i \in s} y_i}{p(s)}$$

and hence it is trivially the best.

Example of a unicluster sampling design

Systematic sampling is a unicluster sampling design.

Consider a systematic sampling scheme of 3 units selected from 12 units. For a systematic sampling scheme, 4 possible samples are as thus

(1, 5, 9), (2, 6, 10), (3, 7, 11) and (4, 8, 12).

The probability of selection in each case is $\frac{1}{4}$, and

$$\pi_i = \frac{1}{4} \qquad \text{for } i = 1,...,12 .$$

For this systematic sample, the only linear unbiased estimator for population total $Y = \sum_{1}^{12} y_i$ is

$$e = \frac{y_s}{p(s)} \quad \text{for } s = 1,2,3,4$$
$$= 4\sum_{i \in s} y_i \ .$$

iii) Basu (1971)

Basu generalized the non existence theorem. He proved that the MVUE does not exist in the class of unbiased estimators.

Theorem

For a non-census design, there does not exist a UMVUE of $\theta(y)$ in the class of any unbiased estimators C.

Proof:

If possible let $T_o(s,y)$ be the UMVUE of the population parameter $\theta(y)$. Since the design p is non-census and the value of $T_o(s,y)$ depends on y_i's for $i \in s$, we can find a vector $y^{(a)} = (a_1,...,a_i,...,a_N)$ for which $T_o(s,y) \neq \theta\left(y^{(a)}\right)$ with $p(s) > 0$.

Consider the following estimator

$$T^*(s,y) = T_o(s,y) - T_o(s,y^{(a)}) + \theta\left(y^{(a)}\right).$$

$T^*(s,y)$ is unbiased for $\theta(y)$ because

$$E_p\left[T^*(s,y^{(a)})\right] = \theta(y) - \theta\left(y^{(a)}\right) + \theta\left(y^{(a)}\right) = \theta(y)$$

Since $T(s,y)$ is the UMVUE for $\theta(y)$, we must have

$$V_p\left[T(s,y)\right] \le V_p\left[T^*(s,y)\right] \qquad \forall y \in R_N .$$

Now for $y = y^{(a)}$

$$V_p\left[T^*(s,y)\right] = V_p\left[T^*(s,y^{(a)})\right] = V_p\left[\theta\left(y^{(a)}\right)\right] = 0 ,$$

while $V_p\left[T^*(s,y^{(a)})\right] > 0$ since we assume $T_o(s,y^{(a)}) \neq \theta\left(y^{(a)}\right)$ with $p(s) > 0$.

Hence the inequality is violated at $y = y^{(a)}$ and the non-existence of a UMVUE for $\theta(y)$ is proved.

3.2 Conclusion

We have seen that an unbiased estimator is not unique, we can derive several unbiased estimators for a fixed sampling design. A natural question is to identify the estimator which is the best. Godambe (1955) first proved that the best estimator does not exist for almost all sampling designs. Therefore one should use his own experience and/or situation to find a suitable estimator. For example, one should construct a πPS sampling design if it is known that the $y_i's$ are approximately equal to the $x_i's$.

Chapter 4

Admissibility

In Chapter 3, we discussed the concepts of unbiasedness and minimum variance unbiased estimators. Following the work by Godambe (1955), Hanurav (1966) and Basu (1971), we noted that there does not exist a minimum variance unbiased estimator when estimating finite population totals except for a unicluster sampling design.

In this chapter, we introduce the concept of admissibility which may guard against an inefficient estimator. Godambe (1960) proved that the Horvitz-Thompson (1952) estimator is found to be admissible in the class of linear unbiased estimators. Godambe and Joshi (1965) extended Godambe's result and proved that the Horvitz-Thompson (1952) estimator is admissible in the class of unbiased estimators. We also discuss the concept of sufficient statistics and explain how one can improve an inadmissible estimator using sufficient statistics and the Rao-Blackwell Theorem.

4.1 Admissible Estimator

An estimator T is said to be admissible in a class C of unbiased estimators under a given sampling design p if there does not exist any other estimator in the class C better than T.

i.e. there does not exist an alternate estimator $T^*(\neq T) \in C_\theta$ for which the following inequalities hold:

i) $V_p(T^*) \leq V_p(T)$ $\forall T^*(\neq T) \in C_\theta \ \ \forall y \in R_N$

ii) $V_p(T^*) < V_p(T)$ for at least one $y \in R_N$.

4.2 Admissibility of Horvitz-Thompson (1952) Estimator

4.2.1 Admissibility in the class of linear homogeneous estimators

Godambe (1960) proved that the Horvitz-Thompson (1952) estimator (t_{HTE}) is admissible in the class of linear homogeneous unbiased estimators.

Theorem 4.1

In the class of linear homogeneous unbiased estimators (C_{lh}), t_{HTE} based on a sampling design p (with $\pi_i > 0 \ \forall i = 1,...,N$), is admissible for a population total Y .

Proof:

The class C_{lh} consists of estimators of the form

$$t(s,y) = \sum_{i \in s} b_{s_i} y_i \in C_{lh} \qquad (4.1.1)$$

where the constants b_{s_i}'s are free from y_i's and subject to satisfying the unbiasedness condition:

$$\sum_{s \ni i} b_{s_i}\, p(s) = 1 \qquad \forall i = 1,\ldots,N\,. \tag{4.1.2}$$

Now $$V_p[t(s,y)] = \sum_{s\in\varphi}\left(\sum_{i\in s} b_{s_i} y_i\right)^2 p(s) - Y^2$$

$$= \sum_{s\in\varphi}\left(\sum_{i\in s} {b_{s_i}}^2 {y_i}^2 + \sum_{i\neq}\sum_{j\in s} b_{s_i} b_{s_j} y_i y_j\right) p(s) - Y^2$$

$$= \sum_{i=1}^{N} {y_i}^2\left(\sum_{s\supset i} {b_{s_i}}^2 p(s) - 1\right) - \sum\sum y_i y_j \left(\sum_{s\supset i,\, j} b_{s_i} b_{s_j}\, p(s) - 1\right).$$

Let $\underline{y}(i)$ = vector $\underline{y}$ whose co-ordinates $y_i = 0$ for $i \neq j = 1,\ldots,N$ and $y_j \neq 0$.

Then

$$V_p[t(s,y^{(j)})] = {y_j}^2\left(\sum_{s\supset j} {b_{sj}}^2 p(s) - 1\right) \geq {y_j}^2\left[\frac{\left(\sum_{s\supset j} b_{sj}\, p(s)\right)^2}{\sum_{s\supset j} p(s)} - 1\right] = {y_j}^2\left(\frac{1}{\pi_i} - 1\right) \tag{4.1.3}$$

(Noting the unbiasedness condition $\sum_{s\supset j} b_{s_i}\, p(s) = 1$).

The equality in (4.1.3) holds if and only if $b_{s_j} = \frac{1}{\pi_j}$, so that

$$V_p[t(s,y)] \geq V_p\left(\sum_{i\in s} \frac{y_i}{\pi_i}\right) \qquad \forall\ y = y(j)\,,\ j = 1,\ldots,N\,. \tag{4.1.4}$$

The inequality in (4.1.4) above is strict if and only if $t(s,y) \neq t_{HTE}$.

There thus cannot be any estimator in C_{lh} better than t_{HTE} when vector $y = y(i)$.

Hence t_{HTE} is admissible in C_{lh}.

4.2.2 Admisibility in the class of unbiased estimators

Godambe and Joshi (1965) extended Godambe's (1960) result further and proved the admissibility of t_{HTE} in the class of unbiased estimators.

Theorem 4.2

Estimator t_{HTE} is admissible in the class C_u of unbiased estimators for a finite population total Y under a sampling design p with $\pi_i > 0 \ \forall i = 1,...,N$.

Proof:

Suppose t_{HTE} is not admissible in the class C_u and there exists an estimator $e(s,y)\ (\neq t_{HTE}) \in C_u$ which is better than t_{HTE}. In this case

$$\text{i)} \qquad V_p[e(s,y)] \leq V_p(t_{HTE}) \qquad \forall y \in R_N \quad (4.2.1)$$

and

$$\text{ii)} \qquad V_p[e(s,y)] < V_p(t_{HTE}) \qquad \text{for at least one } y \in R_N. \quad (4.2.2)$$

The estimator $e(s,y)$ can be written as

$$e(s,y) = t_{HTE} + h(s,y) \qquad (4.2.3)$$

where $h(s,y) = e(s,y) - t_{HTE}$.

Since $e(s,y)$ and $t_{HTE} \in C_u$, (4.2.3) yields

$$E_p[h(s,y)] = \sum_{s\in\varphi} h(s,y)\,p(s) = 0\,.$$

Further (4.2.1) implies that

$$[V_p[h(s,y)] + 2C_p[t_{HTE}, h(s,y)] \le 0 \qquad (4.2.4)$$

where C_p denotes covariance with respect to the sampling design p.

Equation (4.2.4) yields

$$\sum_{s\in\varphi}\{h(s,y)\}^2\,p(s) + 2\sum_{s\in\varphi} h(s,y)\left\{\sum_{i\in s}\frac{y_i}{\pi_i}\right\}p(s) \le 0 \qquad \forall y \in R_N\,. \quad (4.2.5)$$

Let us define $\underline{y}(j)$ = collection of all vectors $\underline{y} = (y_1,....,y_k,...,y_N)$ having j nonzero co-ordinates and $N-j$ zero co-ordinates.

Also $\varphi(j)$ $(\subset \varphi)$ is a collection of samples consisting of units with y values that are non - zero for exactly j units.

Clearly $\quad y(j)\cap y(k) = \varnothing \quad$ for $j \ne k = 1,...,N$;

$\varphi(f)\cap\varphi(g) = \varnothing \quad$ for $f \ne g = 1,...,n$;

$$\bigcup_{j=1}^{N} y(j) = R^N \quad \text{and} \quad \bigcup_{j=1}^{n}\varphi(j) = \varphi\,.$$

Now when $y = y(0) = (0,...,0,...,0)$,

$$\sum_{i\in s}\frac{y_i}{\pi_i} = 0 \qquad \text{for every } s \in \varphi$$

then equation (4.2.5) yields

$$h(s, y) = 0 \qquad \forall s \in \varphi . \tag{4.2.6}$$

Now if $h(s, y) = 0 \quad \forall s \in \varphi$ and $\forall y \in y(j)$, then for any $y \in y(j+1)$ the equations (4.2.4) and (4.2.5) yield

$$\sum_{s \in \varphi(j+1)} h(s, y)\, p(s) = 0 \tag{4.2.7}$$

and

$$\sum_{s \in \varphi(j+1)} \{h(s, y)\}^2 \, p(s) + 2 \sum_{s \in \varphi(j+1)} h(s, y) \left\{ \sum_{i \in s} \frac{y_i}{\pi_i} \right\} p(s) \le 0 . \tag{4.2.8}$$

Now $$\sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} \frac{y_i}{\pi_i}$$ for every $s \in \varphi(j+1)$

and hence (4.2.7) and (4.2.8) give

$$\sum_{s \in \varphi(j+1)} \{h(s, y)\}^2 \, p(s) + 2 \sum_{i=1}^{N} \frac{y_i}{\pi_i} \sum_{s \in \varphi(j+1)} h(s, y)\, p(s) \le 0 \qquad \forall s \in \varphi(j+1)$$

and $\forall y \in y(j+1)$

i.e. $h(s, y) = 0 \qquad \forall s \in \varphi(j+1)$ and $\forall y \in y(j+1)$. (4.2.9)

Now from (4.2.6) and (4.2.9) we see that

$$h(s, y) = 0 \qquad \forall s \in \varphi \text{ and } \forall y \in R_N .$$

Thus there does not exist an estimator, $e(s, y)\,(\neq t_{HTE}) \in C_u$ which is better than t_{HTE}. We thus conclude that t_{HTE} is an admissible estimator in the class C_u.

4.3 Inadmissible Estimators

4.3.1 Definition

An estimator $e(s,y)$ is said to be inadmissible in a class C if there exists an estimator $e^*(s,y)$ $(\in C)$ better than $e(s,y)$.

We can always improve an inadmissible estimator by applying the Rao-Blackwell theorem using sufficient statistics. Such an improvement of an inadmissible estimator is known as Rao-Blackwellization. The technique of Rao-Blackwellization is described as follows:

4.3.2 Sufficient Statistics in Finite Population Sampling

Let $s=(u_{i_1},...,u_{i_n})$ be an ordered sample of size n selected from a population U with probability $p(s)$ using a sampling design p, then $d=(i_k,y_{ik};i_k\in s)$ is the ordered data. Let $\tilde{d}$ =($i_k,y_{ik};i_k\in\tilde{s}$) be the unordered data obtained from the ordered data d. The unordered sample $\tilde{s}$, is obtained by taking the set of distinct units in s and ignoring repetition of units in s.

The values of the parameter $y=(y_1,...,y_N)$ are not known before the survey, so $\Omega_y=R_N$ = N dimensional Euclidean space is considered as the parametric space.

After surveying the sample s, the data d = $(i_k,y_{ik};i_k\in s)$ is collected. From this we get $\tilde{d}$, the unordered data. The data $\tilde{d}$ is said to be consistent with the parameter y_0 = $(y_{10},...,y_{i0},...,y_{N0})$ if $y_{j_k}=y_{j_k0}$ for $j_k\in\tilde{s}$

.i.e if $y_{j_1}=y_{j_10},...,y_{j_\nu}=y_{j_\nu 0}$, ν is the number of distinct units in s.

Once the data d is collected, the values of the $y_i's$ belonging to the unordered sample $\tilde{s}$ are known. Hence the parametric space is now $\Omega_{y\tilde{d}}$, which consists of the vectors $\underline{y}$ with $y_j = y_{j0}$ for $j \in \tilde{s}$.

Example 4.3.1

Consider a population of size N = 4 and $y = (y_1, y_2, y_3, y_4)$.

The parametric space is the four dimensional Euclidean space Ω_y.

Suppose an ordered sample s = (1, 3, 3) is selected.

Surveying s yields y_1 = 5 and y_3 = 10.

Then $\tilde{s}$ =(1, 3) and $\Omega_{y\tilde{d}}$ = (5, -∞ < y_2 < ∞ , 10, -∞ < y_4 < ∞).

NOTE: The details are given by Arnab (2006)

Definition

Let $y_1,...,y_n$ be a random sample with unknown parameter $\theta(Y_1,...,Y_n)$ = θ (say).

The statistic $u = g(Y_1,...,Y_n)$ is sufficient for θ if the conditional distribution of $y_1,...,y_n$ given u is not dependent on θ.

Theorem 4.3

The unordered data $\tilde{d}$ is a sufficient statistic for y.

The detailed proof can be obtained from Cassel, Särndal and Wretman (1977)

4.3.3 Rao-Blackwellization in Finite Population Sampling

An ordered sample s is selected with probability $p(s)$ from a population, d is the corresponding ordered data.

Let $t(d)$ be an unbiased estimator for a parametric function $\theta(y)=\theta$ and $t^*(\tilde{d}) = E_p[t(d) \mid \tilde{d}]$ where $\tilde{d}$ is the unordered data obtained from d.

Theorem 4.4

Estimator $t^*(\tilde{d})$ is an unbiased estimator of θ with $V_p[t^*(\tilde{d})] \leq V_p[t(d)]$

Proof:

$$\theta = E_p[t(d)] = E_p\left(E_p[t(d) \mid \tilde{d}]\right) = E_p[t^*(\tilde{d})]$$

and

$$V_p[t(d)] = E_p\left(V_p[t(d) \mid \tilde{d}]\right) + V_p(E_p[t(d) \mid \tilde{d}])$$

$$= E_p\left(V_p[t(d) \mid \tilde{d}]\right) + V_p[t^*(\tilde{d})]$$

$$\geq V_p[t^*(\tilde{d})] \qquad \text{since } E_p(V_p[t(d) \mid \tilde{d}]) \geq 0.$$

4.3.3.1 Examples

i) SRSWR

In the SRSWR sampling scheme the probability of selection of an ordered sample $s=(u_{i_1}, u_{i_2}, ..., u_{i_n})$ is $p(s)=\frac{1}{N^n}$.

Let $y_{(r)}$ be the value of the character under study y for the population, selected on the rth draw.

Let $\bar{y}(s)=\frac{\sum_{r=1}^{n} y_{(r)}}{n}$, then $\bar{y}(s)$ is an unbiased estimator for the population mean $\bar{Y}$. This estimator is inadmissible since it is based on ordered data, possible repetition. This estimator is not based on a sufficient statistic.

Let $\tilde{s}=(u_{j_1},u_{j_2},....,u_{j_v})$ denote the unordered sample obtained by taking v the set of distinct units $j_1,....,j_v$ ($j_1<....<j_v$) in s.

Theorem 4.5

Let $\bar{y}(\tilde{s}) = \frac{\sum_{i\in\tilde{s}} y_i}{v} = \frac{\sum_{k=1}^{v} y_{j_k}}{v}$ be the sample mean based on the distinct units of s. Then

i) $E[\bar{y}(\tilde{s})] = E[\bar{y}(s)] = Y$

and

ii) $V[\bar{y}(\tilde{s})] \leq V[\bar{y}(s)]$.

Proof:

Let $n_i(s)$ denote the number of times the ith unit appears in s.

If

$$\bar{y}(s)=\frac{1}{n}\sum_{r=1}^{n} y_{(r)} = \frac{1}{n}\sum_{i=1}^{N} n_i(s)y_i$$

then

i) $$E\left[\bar{y}(s)\right] = E\left[\bar{y}(s)/\tilde{s}\right] = \frac{1}{n}\sum_{i=1}^{n} E[n_i(s)y_i \mid \tilde{s}] = n\frac{1}{v}\sum_{i\in\tilde{s}}\frac{y_i}{n}$$

Since for a given $\tilde{s}$, $n_i(s)$ follows a multinomial distribution with $E(n_i(s)/\tilde{s}) = n\frac{1}{v}$.

ii) $$V[\bar{y}(s)] = V\{E[\bar{y}(s)/\tilde{s}]\} + E\{V[\bar{y}(s)/\tilde{s}]\}$$
$$\geq V\{E[\bar{y}(s)/\tilde{s}]\}$$
$$= V[\bar{y}(\tilde{s})]$$

which shows that the sample mean based on distinct units is uniformly better than $\bar{y}$ (s) based on all the units.

ii) PPSWR

Hansen-Hurwitz (1943) Estimator

Let a sample $s = (u_{i_1}, u_{i_2}, ..., u_{i_n})$ of size n be selected from a population by PPSWR method of sampling with p_i denoting the normed size measure ($p_i > 0$) for the ith unit. Let $u_{(r)}$ be the unit selected at the rth draw and $p_{(r)}$ be the corresponding normed size measure. If the ith draw produces the rth unit then

$$u_{(r)} = u_i \quad \text{and} \quad p_{(r)} = p_i .$$

Then the estimator

$$\hat{Y}_{hh} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_{(r)}}{p_{(r)}}$$

is unbiased for the population total Y.

The estimator $\hat{Y}_{hh}$ is known as the Hansen-Hurwitz (1943) estimator.

$\hat{Y}_{hh}$ is an ordered estimator since it depends on the multiplicity of the units selected and the order of the selection of the units in the sample s, hence $\hat{Y}_{hh}$ is inadmissible.

Now writing

$$\hat{Y}_{hh} = \frac{1}{n}\sum_{i\in\tilde{s}} n_i(s)\frac{y_i}{p_i}$$

where $n_i(s)$ = number of times the ith unit occurs in s.

$$E\left(\hat{Y}_{hh}/\tilde{s}\right) = \frac{1}{n}\sum_{i\in\tilde{s}} E\left(n_i(s)/\tilde{s}\right)\frac{y_i}{p_i}$$

$$= \frac{\sum_{i\in\tilde{s}} y_i}{\sum_{i\in\tilde{s}} p_i}.$$

Clearly $\dfrac{\sum_{i\in\tilde{s}} y_i}{\sum_{i\in\tilde{s}} p_i}$ has a smaller variance than $\hat{Y}_{hh}$.

Let $\tilde{s}$ be the unordered sample obtained by taking distinct units of the selected ordered sample s, then applying the Rao-Blackwell Theorem, one can find an improved estimator as shown in the next example.

Example 4.3.2

Let $s_1 = (i, i, j)$ be an ordered sample of size n=3 selected by PPSWR method. The Hansen-Hurwitz (1943) estimator based on s_1 is given by

$$\hat{Y}_{hh}(s_1) = \frac{1}{3}\left(2\frac{y_i}{p_i} + \frac{y_j}{p_j}\right).$$

From the ordered sample s_1, we get the unordered sample $\tilde{s} = (i, j)$ with $i < j$.

The unordered sample $\tilde{s}$ could be realized from any of the following ordered samples s:

$$s_1 = (i,i,j), \quad s_2 = (i,j,i), \quad s_3 = (j,i,i)$$
$$s_4 = (i,j,j), \quad s_5 = (j,i,j), \quad s_6 = (j,j,i).$$

Now since

$$\hat{Y}_{hh}(s_1) = \hat{Y}_{hh}(s_2) = \hat{Y}_{hh}(s_3) = \frac{1}{3}\left(2\frac{y_i}{p_i} + \frac{y_j}{p_j}\right),$$

$$\hat{Y}_{hh}(s_4) = \hat{Y}_{hh}(s_5) = \hat{Y}_{hh}(s_6) = \frac{1}{3}\left(\frac{y_i}{p_i} + 2\frac{y_j}{p_j}\right),$$

$p(s_1) = p(s_2) = p(s_3) = p_i^{\,2} p_j$ and

$p(s_4) = p(s_5) = p(s_6) = p_i p_j^{\,2}$.

We get the following unordered estimator

$$t^* = E\left[\hat{Y}_{hh} / \tilde{s}\right] = \frac{\sum_{k=1}^{6} \hat{Y}_{hh}(s\ ) p(s_k)}{\sum_{k=1}^{6} p(s_k)}$$

$$= \frac{1}{3}\left(\frac{y_i}{p_i} + \frac{y_j}{p_j} + \frac{y_i + y_j}{p_i + p_j}\right).$$

iii) PPSWOR

Suppose on the first draw the ith unit is selected with probability p_i. At the second draw $jth(\neq i)$ unit is selected with probability $p_j = \frac{p_j}{1-p_i}$.

Raj's (1956) Estimator

$$\hat{Y}_{RAJ} = \frac{1}{n}\sum_{i=1}^{n} t(i_r)$$

The above estimator is ordered since it depends on the order of selection of units in the ordered sample.

Consider an ordered sample $s = (i, j)$ of size $n = 2$.

Then

$$t(i) = \frac{y_i}{p_i} \qquad \text{and} \qquad t(j) = y_i + \frac{y_j}{p_j}(1-p_i).$$

So Raj's estimator based on the ordered sample $s = (i, j)$ is

$$\hat{Y}_{RAJ}(i, j) = \frac{1}{2}\left[t_1(i, j) + t_2(i, j)\right]$$

$$= \frac{1}{2}\left[\frac{y_i}{p_i}(1-p_i) + \frac{y_j}{p_j}(1-p_i)\right].$$

However Raj's estimator based on the ordered sample $s^* = (j, i)$ is

$$\hat{Y}_{RAJ}(j, i) = \frac{1}{2}\left[t_1(j, i) + t_2(j, i)\right]$$

$$= \frac{1}{2}\left[\frac{y_j}{p_j}(1-p_j) + \frac{y_i}{p_i}(1-p_j)\right].$$

So $\hat{Y}_{RAJ}(i, j) \neq \hat{Y}_{RAJ}(j, i)$.

Modification of Raj's (1956) Estimator – Murthy (1957)

Murthy's unordered estimator is obtained by taking the weighted average of Raj's estimator with weights proportional to the selection probability of the ordered sample.

So Murthy's estimator based on the ordered samples $s = (i, j)$ or $s^* = (j, i)$ is given by:

$$\hat{Y}_{MUR} = \frac{\hat{Y}_{RAJ}(s)p(s) + \hat{Y}_{RAJ}(s^*)p(s^*)}{p(s) + p(s^*)}$$

$$= \left[\frac{y_i}{p_i}(1-p_i) + \frac{y_j}{p_j}(1-p_j)\right] \Big/ \left(2 - p_i - p_j\right).$$

This is an unordered estimator since it is independent of selection of the order of the sample.

Hence we get the following theorem which states that both Raj's estimator and Murthy's estimator are unbiased estimators of the population total Y and that Murthy's estimator is better than Raj's estimator since it has a smaller variance:

Theorem 4.6

(1) $E[t_{RAJ}] = E[t_{MUR}] = Y$

and

(2) $V[t_{MUR}] \leq V[t_{RAJ}]$.

The proof of the above result can be found in Murthy (1957).

4.2 Conclusion

The criterion of admissibility, like sufficiency, does not single out a unique estimator. Many traditional estimators in survey sampling have been shown to be admissible. Hanurav (1965, 68) proposed the criteria of hyperadmissibility, a stronger form of admissibility. The proposed criteria of hyperadmissibility singles out one estimator, the Horvitz-Thompson (1952) estimator, as the unique hyperadmissible estimator in the class of linear homogeneous unbiased estimators and also in the class of unbiased estimators. We have not discussed the concept of hyperadmissibility in this thesis.

We should try to avoid the use of inadmissible estimators. As a rule of thumb, to find an admissible estimator, we must not choose an estimator which is:

i) based on the order of selection of units and/or repetition and

ii) not based on a sufficient statistic.

However, we use inadmissible estimators in various situations for their simplicity and elegant expressions of variance. For example, sample mean based on SRSWR and the Hansen-Hurwitz (1943) estimator based on PPSWR sampling.

It is also important to note that the Rao-Hartley-Cochran estimator (discussed in Chapter 6) is used extensively for its simplicity even though it is known to be inadmissible.

Chapter 5

Superpopulation Model

In the model based approach, also known as the prediction approach, it is assumed that the population y-values are random and obey a model (known as superpopulation model) and that the model distribution leads to valid inference referring to a particular sample that has been drawn irrespective of the sampling design. Model based inference, in large samples however, are sensitive to model misspecifications as illustrated by Hansen, Madow and Tepping (1983).

We also describe the model-design or model assisted approach which is a hybrid of the design based and model based approach. In this approach, inference is based on the sampling design as well as superpopulation models. Details are given by Rao (1994) as well as in Cassel, Särndal and Wretman (1977).

In this chapter, we present optimum estimators of finite population characteristics using model based and design based approaches. It is found that the Horvitz-Thompson (1952) estimator becomes optimal under various superpopulation models providing an appropriate sampling design is used.

5.1 Superpopulation Model

In the previous chapters we discussed design based inference where the population vector $\underline{y} = (y_1,...,y_N)$ was a fixed point in the N-dimensional Euclidean space. In that case, we found that there does not exist a uniformly minimum variance unbiased estimator in the class of unbiased estimators for estimating the population total Y.

In this chapter we consider the population vector $\underline{y}$ as a realization of a random variable $Y = (Y_1,...,Y_N)$ and its distribution will be denoted by ξ.

The probability distribution ξ may depend on a parameter θ, which is generally unknown and belongs to a certain known parameter space Ω_θ. Such a probability distribution ξ is known as a superpopulation model. In most situations, the distribution ξ is related to a fixed auxiliary variable $x = (x_1,...,x_N)$ whose elements are assumed to be known and positive.

Example 5.1

Let us consider the exam marks of 125 first year statistics students at a certain university in 2006.

The vector $y = (y_1, y_2,..., y_{125})$ is the exam mark for the students, i.e. y_1 = exam mark for student 1, y_2 = exam mark for student 2,... etc.

If we consider the students for different years, then the vector $\underline{y}$ will take on different values. Here we consider a distribution of $\underline{y}$, which will be called a superpopulation model.

5.2 Definitions

For a superpopulation model ξ and sampling design p, the expectation, variance and co-variance operators are denoted by E_ξ, V_ξ, C_ξ and E_p, V_p, C_p respectively.

5.2.1 Design Unbiased (*p* - unbiased) Estimator

An estimator t is said to be *design unbiased* for total Y if and only if

$$E_p(t) = Y \qquad \forall y \in R_N .$$

The class of *p*-unbiased estimators will be denoted by C_p.

5.2.2 Model Unbiased (ξ - unbiased) Estimator

An estimator t is said to be model unbiased if and only if

$$E_\xi(t) = E_\xi(Y) \qquad \forall \theta \in \Omega_\xi .$$

The class of ξ-unbiased estimators will be denoted by C_ξ.

5.2.3 Model-Design Unbiased (*p*ξ - unbiased) Estimator

An estimator t is said to be a model-design unbiased estimator if and only if

$$E_\xi E_p(t) = E_\xi(Y) \qquad \forall \theta \in \Omega_\xi .$$

The class of $p\xi$-unbiased estimators will be denoted by $C_{p\xi}$.

If an estimator is design and model unbiased then it is model-design unbiased i.e. the class of model-design unbiased estimators contains both the class of design unbiased estimators C_p and the class of model unbiased estimators C_ξ .

5.2.4 Non-informative Sampling Design

A sampling design is said to be a non-informative sampling design if and only if the selection of a sample does not depend on the study variable $y_i's$ i.e. the sampling design is non-sequential.

For a non-informative sampling design E_ξ and E_p are commutative i.e.

$$E_\xi E_p(t) = E_p E_\xi(t) .$$

5.2.5 Optimal Estimator

An estimator t_0 belonging to a certain class of estimators C, is said to be an optimal estimator (or optimal) for estimating Y under a given superpopulation model ξ and a sampling design p if

$$E_\xi E_p(t_0 - Y)^2 \leq E_\xi E_p(t - Y)^2 \qquad \forall t(\neq t_0) \in C, \theta \in \Omega_\theta$$

and the inequality is strict for some $\theta \in \Omega_\theta$.

5.2.6 Optimal Strategies

A sampling strategy $h(p,t)$ with $p \in P, t \in C$, is a combination of sampling design p and estimator t, based on a sample selected using the design p.

Let H be a class of strategies $h=(p,t)$ with $p \in P, t \in C$, then the strategy $h_0=(p_0,t_0) \in H$ is said to be optimal in H

if

$$E_\xi E_{p_0}(t_0 - Y)^2 \leq E_\xi E_p (t - Y)^2 \qquad \forall h(\neq h_0) \in H, \theta \in \Omega_\theta$$

and the inequality is strict for some $\theta \in \Omega_\theta$.

5.3 Inference under Model-based approach

Suppose we have collected the data $d = \{(i, y_i), i \in s\}$ where the values of y_i in the sample s have been recorded. In the prediction approach, the statistician is to predict the unobserved values of y_i for $i \notin s$ i.e. $i \in U - s$, U being the finite population. This is done by assuming a superpopulation model where the actual values $\underline{y} = (y_1,...,y_N)$ are one of the realizations of the random variables $\underline{Y} = (Y_1,...,Y_N)$. The joint probability distribution of Y supplies the link between the observed y_i's $i \in s$ and the unobserved y_i's $i \notin s$.

The details are given by Royall (1970), Cassel, Särndal and Wretman (1977), Chaudhuri and Stenger (1992), Lohr (1999) and Valliant, Dorfman and Royall (2000) amongst others.

5.3.1 Estimation of Population Total Y

Here we assume the following superpopulation model

Model ξ : $E_\xi(Y_i) = \beta x_i$, $V_\xi(Y_i) = \sigma_i^2$ and $C_\xi(Y_i, Y_j) = 0$ (5.3.1)

where E_ξ, V_ξ and C_ξ denote the expectation, variance and covariance with respect to the model ξ,

$x_i's$ are known, positive auxiliary variable,

β is a model parameter and

$\sigma_i^2 = \sigma^2(x_i) =$ particular function of x_i only.

The population total Y can be written as

$$Y = \sum_{i \in s} y_i + \sum_{i \notin s} y_i \ . \tag{5.3.2}$$

The quantity $\sum_{i \in s} y_i$ is known because y_i, $i \in s$ has been observed. We need to predict the unobserved quantity $\sum_{i \notin s} y_i$ using the superpopulation model ξ.

Consider the conditional expectation given data $d = \{(i, y_i), i \in s\}$ viz.

$$E_\xi\left(\sum_{i \notin s} y_i / d\right) = \beta \sum_{i \notin s} x_i \ . \tag{5.3.3}$$

Now the quantity $\sum_{i \notin s} x_i$ in (5.3.3) is known and we predict β through the data d collected.

We may use the following linear function for prediction of β viz.

$$\hat{\beta} = \sum_{i \in s} d_{i_s} y_i$$

where $d_{i_s}'s$ are known constants independent of $y_i's$.

Now replacing $\sum_{i \notin s} y_i$ by its predicted value $\hat{\beta} \sum_{i \notin s} x_i$ in (5.3.2), we get the estimator

$$t = \sum_{i \in s} y_i + \hat{\beta} \sum_{i \notin s} x_i \qquad (5.3.4)$$

where $\hat{\beta} = \sum_{i \in s} d_{i_s} y_i$.

The estimator t in equation (5.3.4) is called a predictor for Y.

Definition 1: The predictor t is called a linear model unbiased predictor for Y if

$$E_\xi(y) = E_\xi(Y) = \beta X \qquad (5.3.5)$$

where $X = \sum_{i=1}^{N} x_i$

i.e. $$\beta \sum_{i \in s} x_i + \left(E_\xi(\hat{\beta})\right) \sum_{i \notin s} x_i = \beta X \,. \qquad (5.3.6)$$

The equation (5.3.6) gives the condition for model unbiasedness of t as follows:

$$E_\xi\left(\hat{\beta}\right) = \beta \,. \qquad (5.3.7)$$

For the linear model unbiased predictor t given in equation (5.3.4), we may choose a loss function

$$M(t) = V_\xi(t - Y) = E_\xi\left[(t - Y) - E_\xi(t - Y)\right]^2 .$$

Definition 2: Best linear optimum predictor

Let C_ξ be the class of all linear model unbiased predictors $t = \sum_{i \in s} d_{s_i} y_i$

satisfying

$$E_\xi(y) = E_\xi(Y) = \beta X \, .$$

The predictor t_0 will be called the best linear unbiased predictor (BLUP) for Y if

$$V_\xi(t_0 - Y) \le V_\xi(t - Y)^2 \qquad \forall t(\neq t_0) \in C_\xi \, .$$

Theorem

Under the superpopulation model

$$Y_i = \beta x_i + \in_i \qquad (5.3.8)$$

where

$E_\xi(Y_i) = 0$, $V_\xi(Y_i) = \sigma_i^2$ and $C_\xi(Y_i, Y_j) = 0$,

the optimum linear predictor is

$$t_{BLUP} = \sum_{i \in s} y_i + \hat{\beta}_0 \sum_{i \notin s} x_i \qquad (5.3.9)$$

for $\hat{\beta}_0 = \dfrac{\sum_{i \in s} \frac{y_i x_i}{\sigma_i^2}}{\sum_{i \in s} \frac{x_i^2}{\sigma_i^2}} \sum_{i \notin s} x_i$.

Proof:

$$V_\xi(t-Y) = V_\xi(t) + V_\xi(Y) - 2C_\xi(t,Y)$$

Now $V_\xi(t) = V_\xi\left(\sum_{i\in s} d_{i_s} y_i\right) = \sum_{i\in s} d_{i_s}^2 \sigma_i^2$,

$V_\xi(Y) = \sum_{i=1}^{N} \sigma_i^2$ and

$C_\xi(t,Y) = C_\xi\left(\sum_{i\in s} d_{i_s} y_i, \sum_{i=1}^{N} y_i\right) = \sum_{i\in s} d_{i_s} \sigma_i^2$.

To minimize $V_\xi(t-Y)$ subject to the condition $E_\xi(t) = E_\xi(Y) = \beta X$, we construct

$$\begin{aligned}\phi &= V_\xi(t-Y) - \lambda\left[E_\xi(t) - E_\xi(Y)\right] \\ &= \sum_{i\in s} d_{i_s}{}^r \sigma_i^2 + \sum_{i=1}^{N} \sigma_i^2 - 2\sum_{i\in s} d_{i_s} \sigma_i^2 - \lambda\left(\sum_{i\in s} d_{i_s} x_i - X\right).\end{aligned}$$

Now $$\frac{\partial \phi}{\partial d_{i_s}} = 0$$

$$\Rightarrow 2d_{i_s}\sigma_i^2 - 2\sigma_i^2 - \lambda x_i = 0$$

i.e. $$d_{i_s} = 1 + \frac{\lambda}{2}\frac{x_i}{\sigma_i^2}$$

$$\sum_{i\in s} d_{i_s} x_i = X$$

$$\Rightarrow X = \sum_{i\in s} d_{i_s} x_i = \sum_{i\in s} x_i + \frac{\lambda}{2}\sum_{i\in s}\frac{x_i^2}{\sigma_i^2}$$

i.e. $$\frac{\lambda}{2} = \frac{X - \sum_{i\in s} x_i}{\sum_{i\in s}\frac{x_i^2}{\sigma_i^2}} = \frac{\sum_{i\notin s} x_i}{\sum_{i\in s}\frac{x_i^2}{\sigma_i^2}} \ .$$

Here $d_{i_s} = 1 + \left(\frac{\sum_{i \notin s} x_i}{\sum_{i \in s} \frac{x_i^2}{\sigma_i^2}} \right) \frac{x_i}{\sigma_i^2} = d_{i_0}(s)$ (say)

Here the BLUP is

$$t_{BLUP} = \sum_{i \in s} d_{i_0}(s) y_i = \sum_{i \in s} y_i + \left(\frac{\sum_{i \notin s} \frac{x_i y_i}{\sigma_i^2}}{\sum_{i \in s} \frac{x_i^2}{\sigma_i^2}} \right) \sum_{i \notin s} x_i$$

$$= \sum_{i \in s} y_i + \hat{\beta}_0 \sum_{i \notin s} x_i \ .$$

Corollary 1

If $\sigma_i^2 = \sigma^2 x_i^2$, then $\hat{\beta}_0 = \frac{1}{n} \sum \frac{y_i}{x_i}$ and the BLUP is

$$t_0 = \left(\sum_{i \in s} y_i \right) + \left(\frac{1}{n} \sum_{i \in s} \frac{y_i}{x_i} \right) \sum_{i \notin s} x_i \ .$$

Corollary 2

Let $\sigma_i^2 = \sigma^2 x_i$, then $\hat{\beta}_0 = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i}$ and the optimum BLUP reduces to the ratio estimator as follows

$$t_R = \sum_{i \in s} y_i + \left(\frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i} \right) \left(X - \sum_{i \notin s} x_i \right) = t \left(\frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i} \right) X = \frac{y_s}{x_s} X$$

where $y_s = \sum_{i \in s} y_i$ and $x_s = \sum_{i \in s} x_i$.

5.3.2 Purposive Sampling

The best linear unbiased predictor for the model ξ with $\sigma_i^2 = \sigma^2 x_i$ is given by

$$t_R = \frac{y_s}{x_s} X \ .$$

The magnitude of

$$M(t_R) = V_\xi (t - Y) = V_\xi \left(X \frac{y_s}{x_s} - \sum_{i \notin s} y_i \right)$$

$$= V_\xi \left[x_s{}' \frac{y_s}{x_s} - \sum_{i \notin s} x_i \right] = \frac{x_s{}'^2}{x_s^2} \sum_{i \in s} \sigma_i^2 + \sum_{i \notin s} \sigma_i^2 \ .$$

Now noting $\sigma_i^2 = \sigma^2 x_i$, we get

$$M(t_R) = \sigma^2 X \left(\frac{X}{x_s - 1} \right).$$

The value $M(t_R)$ attains a minimum when $x_s = \sum_{i \in s} x_i$ is the maximum. So the value of $M(t_R)$ attains a minimum value for the sample s if we choose the units with the largest $x_i{}'s$ to constitute the sample.

Now if we choose the optimum sampling design as one which minimizes

$$E_p E_\xi (t_R - Y),$$

then we find the optimal strategy constitutes the estimator t_R and sampling design p_0 which selects the sample s_0 with probability 1. The sampling design p_0 is clearly a purposive sampling design which selects the sample s_0 with probability 1.

5.3.3 Balancing and Robustness

In practice we will never be sure as to which model is appropriate in a given situation. Suppose that model ξ given in equation (5.5.1) is considered adequate and one thinks of adopting the optimal strategy (p_0, t_R) for which

$$t_R = X\frac{\bar{y}}{\bar{x}}$$

and

$$V\xi(t_R - Y) = M_0 = \frac{N^2\left(1-\frac{N}{n}\right)}{n}\frac{\bar{X}\bar{x}_R}{\bar{x}}\sigma^2.$$

We want to examine what happens to the performance of the strategy if the correct model is given by

Model ξ^*: $E_{\xi^*} = \alpha + \beta x_i$.

Under this model $E_{\xi^*}(t_R) = N\alpha\,\bar{X}/\bar{x} + \beta X$ and thus t_R has the following bias

$$B_{\xi^*}(t_R) = E_{\xi}(t_R - Y) = N\alpha\left(\bar{X}/\dot{x} - 1\right).$$

This bias disappears if and only if $\bar{x} = \bar{X}$. Therefore instead of using the design p_0 which is optimal under model ξ, one would use the design p^* where $\bar{x} = \bar{X}$, then t_R which is model unbiased under model ξ is also model unbiased under model ξ^*. A sample for which $\bar{x} = \bar{X}$ is called a *balanced sample* and a design which prescribes choosing a balanced sample with probability 1 is called a *balanced design*. So, based on a balanced sample, t_R is *robust* in respect of model bias.

A balanced design may however not be available if for example there exists no sample of a given size with $\bar{x} = \bar{X}$.

5.4 Optimal Design-unbiased estimators

Here we will show the existence of an optimal estimator in the class of unbiased estimators C_p with the following superpopulation model M1 as defined below.

5.4.1 Model M1

The $y_i's$ are independently distributed with mean $E_{M1}(y_i)=\mu_i$ and variance $V_{M1}(y_i)=\sigma_i^2$ for $i=1,...,N$, where the $\mu_i's$ are known and the $\sigma_i's$ are unknown.

Theorem 5.4.1 (Godambe & Joshi (1965))

Under the model M1 and a given sampling design p with $\pi_i>0 \ \ \forall i=1,...,N$, the expected variance of an unbiased estimator t ($\in C_p$) of Y satisfies the following inequality:

$$E_{M1}V_p(t) \geq \sum_{i=1}^{N}\sigma_i^2\left(\frac{1}{\pi_i}-1\right) = E_{M1}V_p(t_\mu) \qquad (5.4.1)$$

where $t_\mu = \sum_{i=1}^{N}\frac{y_i-\mu_i}{\pi_i}+\mu$ with $\mu=\sum_{i=1}^{N}\mu_i$.

Proof:

$$\begin{aligned}
E_{M1}V_p(t) &= E_{M1}\left[E_p(t^2)-Y^2\right] \\
&= E_pE_{M1}(t^2)-V_{M1}(Y)-\left[E_{M1}(Y)\right]^2 \\
&= E_p\left[V_{M1}(t)\right]+E_p\left[E_{M1}(t)\right]^2-V_{M1}(Y)-\left[E_{M1}(Y)\right]^2 \\
&= E_p\left[V_{M1}(t)\right]+E_p\left[E_{M1}(t)-E_{M1}(Y)\right]^2-V_{M1}(Y) \qquad (5.4.2)
\end{aligned}$$

We let $t(s)$ be the value of the estimator t based on the sample s, selected with probability $p(s)$.

Let us write

$$t(s,y) = t_{HTE}(s,y) + h(s,y)$$

where $t_{HTE}(s,y) = \sum_{i=1}^{N} \frac{y_i}{\pi_i} I_{si}$ is the Horvitz-Thompson estimator.(Horvitz-Thompson (1952))

and $h(s,y)$ is a function of the $y_i's$ for $i \in s$ only.

Since $t(s,y)$ is unbiased for Y, we get

$$\sum_s t(s,y)p(s) = \sum_s t_{HTE}(s,y)p(s) + h(s,y)p(s) = Y$$

which implies that $\sum_s h(s,y)p(s) = 0$. (5.4.3)

Further $\sum_s h(s,y)p(s) = 0$ yields

$$\sum_{s \supset i} h(s,y)p(s) + \sum_{s \not\ni i} h(s,y)p(s) = 0 \qquad (5.4.4)$$

where $\sum_{s \not\ni i}$ is the sum over those samples which do not contain the unit i.

Then we have

$$E_p V_{M1}(t) = \sum_s \left[V_{M1}\{t_{HTE}(s,y)\} + V_{M1}\{h(s,y)\} + 2C_{M1}\{t_{HTE}(s,y), h(s,y)\}\right]p(s). \qquad (5.4.5)$$

Now $$\sum_s V_{M1}\{t_{HTE}(s,y)\}p(s) = \sum_s \sum_{i=1}^{N} \frac{\sigma_i^2}{\pi_i^2} I_{si}\, p(s) = \sum_{i=1}^{N} \frac{\sigma_i^2}{\pi_i}.$$

And $$\sum_s C_{M1}\{t_{HTE}(s,y), h(s,y)\}p(s) = \sum_s p(s) \sum_{i=1}^{N} I_{si} \frac{E_{M1}(y_i - \mu_i)}{\pi_i} h(s,y)$$

$$= \sum_{i=1}^{N} E_{M1}\left[\frac{(y_i - \mu_i)}{\pi_i} \sum_s I_{si} h(s,y)\, p(s)\right]$$

$$= \sum_{i=1}^{N} E_{M1}\left[\frac{(y_i - \mu_i)}{\pi_i} \sum_{s \supset i} h(s,y)\, p(s)\right]$$

$$= -\sum_{i=1}^{N} E_{M1}\left[\frac{(y_i - \mu_i)}{\pi_i} \sum_{s \not\supset i} h(s,y)\, p(s)\right] \qquad \text{(using 5.4.3)}$$

$$= -\sum_{i=1}^{N}\left[E_{M1}\frac{(y_i - \mu_i)}{\pi_i} E_{M1}\left(\sum_{s \not\supset i} h(s,y)\, p(s)\right)\right]$$

$$= 0 \qquad (5.4.6)$$

(as y_i's are independent).

Finally putting (5.4.5) and (5.4.6) into (5.4.2), we get

$$E_{M1}V_p(t) = \sum_{i=1}^{N} \frac{\sigma_i^2}{\pi_i} + E_p\left[V_{M1}\{h(s,y)\}\right] + E_p\left[E_{M1}(t) - E_{M1}(Y)\right]^2 - V_{M1}(Y)$$

$$\geq \sum_{i=1}^{N} \sigma_i^2 \left(\frac{1}{\pi_i} - 1\right).$$

Note:

$E_{M1}V_p(t)$ attains its lower bound (5.4.2) when

i) $E_p\left[V_{M1}\{h(s,y)\}\right] = 0$ and

ii) $E_{M1}(t) - E_{M1}(Y) = 0$. (5.4.7)

These conditions (5.3.7) are satisfied as per Chaudhuri and Stenger (1992) when

$$t = t_{\mu} = \sum_{i=1}^{N} \frac{y_i - \mu_i}{\pi_i} I_{si} + \mu \qquad \text{with } \mu = \sum_{i=1}^{N} \mu_i \, .$$

The estimator

$$t_{\mu} = \sum_{i=1}^{N} \frac{y_i - \mu_i}{\pi_i} I_{si} + \mu$$

is known as the generalized difference estimator.

Consider the model M1 with $\mu_i = \beta x_i$, where β is an unknown positive quantity and x_i is the value of the auxiliary characteristics x for the ith unit which is known and positive for every $i = 1,...,N$.

Let P denote the class of fixed effective size (n) sampling designs and $p_x (\in P_n)$ be a πps design satisfying

$$\pi_i = np_i \qquad \text{for every } i = 1,...,N$$

with $p_i = x_i / X$, $X = \sum_{i=1}^{N} x_i$.

Then $E_{\xi} V_p(t)$ attains the lower bound of (5.3.1) when

$$t = t_{HTE} = \sum_{i=1}^{N} \frac{y_i}{np_i} I_{si} \, .$$

The following theorems were obtained from Godambe and Joshi (1965) and Cassel, Särdal and Wretman (1977).

Theorem 5.4.2

Under the model M1 with $\mu_i = \beta x_i$ and $p \in P_n$

$$E_{M1}V_p(t) \geq \sum_{i=1}^{N} \sigma_i^2 \left(\frac{1}{np_i} - 1 \right) = E_{M1}V_p(t_{HTE}) \qquad \forall t \in C_u . \qquad (5.4.8)$$

Minimizing $\sum_{i=1}^{N} \sigma_i^2 \left(\frac{1}{\pi_i} - 1 \right)$, the right hand side of (5.4.1) subject to $\sum_{i=1}^{N} \pi_i = n$, yields

$$\pi_i = n\sigma_i / \sum_{i=1}^{N} \sigma_i = \pi_i(\sigma)$$

and the corresponding minimum value of

$$\sum_{i=1}^{N} \sigma_i^2 \left(\frac{1}{\pi_i} - 1 \right) = \frac{1}{n} \left(\sum_{i=1}^{N} \sigma_i \right)^2 - \sum_{i=1}^{N} \sigma_i^2 .$$

Let $p_{\sigma,\mu}$ be a fixed effective size sampling design with

$$\pi_i = \pi_i(\sigma) \quad \text{and} \quad \sum_{i=1}^{N} \frac{\mu_i}{\pi_i(\sigma)} I_{si} = \sum_{i=1}^{N} \mu_i \quad \text{for every s with } p(s) > 0,$$

t_μ reduces to $\sum_{i=1}^{N} \frac{y_i}{\pi_i(\sigma)} I_{si}$ as

$$t_\mu = \sum_{i=1}^{N} \frac{y_i - \mu_i}{\pi_i(\sigma)} I_{si} + \mu$$

$$= \sum_{i=1}^{N} \frac{y_i}{\pi_i(\sigma)} I_{si} .$$

Then the expected variance of the Horvitz-Thompson estimator (t_{HTE}) based on sampling design $p_{\sigma,\mu}$ attains the lower bound given above. The following theorem shows how a sampling strategy based on the design $p_{\sigma,\mu}$ and the Horvitz-Thompson estimator is optimal.

Theorem 5.4.3

Under the model M1, $h_o = (p_{\sigma,\mu}, t_{HTE})$, where $t_{HTE} = \sum_{i=1}^{N} \frac{y_i}{\pi_i} I_{si} = x \sum_{i=1}^{N} \frac{y_i}{nx_i}$, is optimal in the class of strategies $H = (p,t)$ with $p \in P_n, t \in C_u$ i.e.

$$E_\xi V_p(t) \geq \frac{1}{n}\left(\sum_{i=1}^{N} \sigma_i\right)^2 - \sum_{i=1}^{N} \sigma_i^2 = E_\xi V_{p_{\sigma,\mu}}(t_{HTE}) \qquad \forall p \in P_n, t \in C_u . \qquad (5.4.9)$$

Another optimal strategy can be found if $\mu_i = \beta x_i$ and $\sigma_i^2 = \sigma^2 x_i^2$ ($\sigma > 0$), as $p_{\sigma,\mu}$ reduces to a p_x design, with

$$\pi_i = n\frac{x_i}{x} = np_i \text{ and } t_{HTE} = x\sum_{i=1}^{N} \frac{y_i}{nx_i} I_{si} ,$$

So we get the following theorem which states that the new strategy based on the design p_x and the estimator t_{HTE} is an optimal strategy.

Theorem 5.4.4

Under the model M1, with $\mu_i = \beta x_i$ and $\sigma_i^2 = \sigma^2 x_i^2$, $h_x = (p_x, t_{HTE})$ is the optimal strategy in the class of strategies $H = (p,t)$ with $p \in P_n, t \in C_u$. i.e

$$E_\xi V_p(t) \geq \sigma^2\left(\frac{X^2}{n} - \sum_{i=1}^{N} x_i^2\right) = E_\xi V_{p_{\sigma,\mu}}(t_{HTE}) \qquad \forall p \in P_n, t \in C_u . \qquad (5.4.10)$$

Finally, the last case that we consider for model M1 is when x_i = 1 for $i = 1,...,N$. We get $\mu_i = \beta x_i$ and $\sigma_i^2 = \sigma^2$ so that design p_x reduces to a sampling design p_0 with $\pi_i = \frac{n}{N} = \pi_0$ and $t_{HTE} = N\bar{y}_s$, where $\bar{y}_s = \sum_{i \in s} y_i / n$. So using the new design we get the following optimal strategy:

Theorem 5.4.5

Under the model M1, with $\mu_i = \beta$ and $\sigma_i^2 = \sigma^2$, $h_0 = (p_0, \bar{y}_s)$ is the optimal strategy in the class of strategies $H = (p,t)$ with $p \in P_n, t \in C_u$. i.e

$$E_\xi V_p(t) \geq \sigma^2 N\left(\frac{N}{n} - 1\right) = E_\xi V_{p_{\sigma,\mu}}(\bar{y}_s) \qquad \forall p \in P_n, t \in C_u . \qquad (5.4.11)$$

5.4.2 Model M2

The next model that we consider is model M2 where

$$E_{M2}(y_i) = \mu_i \ \ (-\infty < \mu_i < \infty), \ V_{M2}(y_i) = \sigma_i^2 \ \ (>0)$$

and

$$C_{M2}(y_i, y_j) = \rho \sigma_i \sigma_j \ \ (-1 < \rho < 1).$$

This model was considered by Cassel, Särdal and Wretman (1977) and Chaudhuri and Stenger (1992) amongst others.

We will first find an optimal estimator and then a few optimal strategies.

Let C_{lu} be the class of linear p-unbiased estimators of the population total Y consisting of estimators of the form

$$t = a_s + \sum_{i \in s} b_{si} y_i$$

where a_s and the $b_{si}'s$ are constants free of the $y_i's$ and satisfy the p-unbiasedness conditions

i) $\sum_s a_s p(s) = 0$

and

ii) $\sum_{s \supset i} b_{si} p(s) = 1 \qquad \forall i = 1,...,N$. (5.4.12)

Now we will find an optimal estimator t_{lo}.

Using equation (5.4.2)

$$E_{M2}V_p(t) = E_p\left[V_{M2}(t)\right] + E_p\left[E_{M2}(t) - E_{M2}(Y)\right]^2 - V_{M2}(Y)$$

$$\geq E_p\left[V_{M2}(t)\right] - V_{M2}(Y) \qquad (5.4.13)$$

$$E_{M2}V_p(t) = E_p\left[\sum_{i \in s} b_{si}^2 \sigma_i^2 + \rho \sum_{i \neq} \sum_{j \in s} b_{si} b_{sj} \sigma_i \sigma_j\right]$$

$$= E_p\left[\left(\sum_{i \in s} b_{si}\sigma_i\right)^2 - (1-\rho)\left\{\left(\sum_{i \in s} b_{si}\sigma_i\right)^2 - \left(\sum_{i \in s} b_{si}^2 \sigma_i^2\right)\right\}\right]$$

$$= \sum_s p(s)\left(\sum_{i \in s} b_{si}\sigma_i\right)^2 - (1-\rho)A \qquad (5.4.14)$$

where $A = \sum_s p(s)\left\{\left(\sum_{i \in s} b_{si}\sigma_i\right)^2 - \left(\sum_{i \in s} b_{si}^2 \sigma_i^2\right)\right\}$.

Now let us maximize A subject to the following condition

$$\sum_{i=1}^{N}\sigma_i\sum_{s\supset i}b_{si}\,p(s)=\sum_{i=1}^{N}\sigma_i\ . \tag{5.4.15}$$

Whenever the $b_{si}'s$ satisfy condition (ii) of equation (5.4.12), they satisfy the condition (5.4.15) above. The converse is not true.

To maximize A subject to the condition (5.4.15), consider the following function ϕ with λ as a Lagrange multiplier:

$$\phi=\sum_{s}p(s)\left\{\left(\sum_{i\in s}b_{si}\sigma_i\right)^2-\left(\sum_{i\in s}b_{si}^2\sigma_i^2\right)\right\}-2\lambda\left\{\sum_{i=1}^{N}\sigma_i\sum_{s\supset i}b_{si}\,p(s)-\sum_{i=1}^{N}\sigma_i\right\}.$$

Differentiating the above function with respect to b_{si} and setting it to equal zero $\left(\frac{\partial\phi}{\partial b_{si}}=0\right)$, we get

$$\sum_{i\in s}b_{si}\sigma_i-\sigma_i b_{si}=\lambda\ . \tag{5.4.16}$$

Summing equation (5.4.15) over $i\in s$ and noting that $\sum_{i\in s}=n$, the sample size, we get

$$\sum_{i\in s}b_{si}\sigma_i=\frac{n\lambda}{n-1}\ . \tag{5.4.17}$$

Multiplying equation (5.4.16) by $p(s)$ and summing over all possible samples yields

$$\sum_{s}p(s)\sum_{i\in s}b_{si}\sigma_i=\frac{n\lambda}{n-1}\ . \tag{5.4.18}$$

Using equation (5.3.15), we get

$$\lambda = \frac{n-1}{n}\sum_{i=1}^{N}\sigma_i \ . \qquad (5.4.19)$$

Substituting equation (5.4.19) into equation (5.4.16) and using equation (5.4.17), we get the optimum values of $b_{si}'s$ which maximize A as

$$b_{si} = b_{si0} = \frac{1}{np_i(\sigma)} \qquad \text{with} \quad p_i(\sigma) = \frac{\sigma_i}{\sum_{i=1}^{N}\sigma_i} . \qquad (5.4.20)$$

Hence

$$A \le \frac{n-1}{n}\left(\sum_{i=1}^{N}\sigma_i\right)^2 . \qquad (5.4.21)$$

The condition (5.4.15) yields,

$$\sum_s p(s)\left(\sum_{i\in s} b_{si}\sigma_i\right)^2 \ge \frac{\left\{\sum_s p(s)\left(\sum_{i\in s} b_{si}\sigma_i\right)^2\right\}}{\sum_s p(s)} = \left(\sum_{i=1}^{N}\sigma_i\right)^2 . \qquad (5.4.22)$$

The equality of equation (5.4.22) holds when $b_{si} = b_{si0} = \frac{1}{np_i(\sigma)}$.

Further with $b_{si} = b_{si0}$, $E_{M2}(t)$ is equal to $E_{M2}(Y)$ if

$$a_s = a_{s0} = \sum_{i=1}^{N}\mu_i - \sum_{i\in s} b_{si0}\mu_i = \mu - \sum_{i\in s}\frac{\mu_i}{np_i(\sigma)} . \qquad (5.4.23)$$

Thus under condition (5.4.15) for any design $p \in P_n$, from equations (5.4.13), (5.4.14), (5.4.21), (5.4.22) and (5.4.23), we get

$$E_{M2}V_p(t) \geq E_p\left[V_{M2}(t)\right] - V_{M2}(Y)$$

$$= (1-\rho)\left\{\frac{\left(\sum_{i=1}^{N}\sigma_i\right)^2}{n} - \left(\sum_{i=1}^{N}\sigma_i\right)^2\right\}$$

$$= E_{M2}V_p(t_{lo}) \qquad (5.4.24)$$

where $t_{lo} = \sum_{i\in s}\frac{y_i - \mu_i}{np_i(\sigma)} + \mu$ and it is an optimal estimator.

The estimator t_{lo} becomes p-unbiased if $\sum_{s\supset i} b_{si0}\,p(s) = 1$, i.e. $\pi_i = np_i$.

Cassel, Särndal and Wretman (1977) showed that if we let $p_{\pi\sigma}$ be the fixed sampling design with inclusion probability $\pi_i = np_i$, a strategy based on this design and the estimator t_{lo} will be the optimal strategy. So we have the following theorem:

Theorem 5.4.6

Under the model M2, $h_{lo} = (p_{\pi\sigma}, t_{lo})$ is optimal in the class of strategies $H = (p,t)$ with $p \in P_n$ and $t \in C_{lu}$ .i.e.

$$E_{M2}V_p(t) \geq E_p\left[V_{M2}(t)\right] - V_{M2}(Y) = (1-\rho)\left\{\frac{\left(\sum_{i=1}^{N}\sigma_i\right)^2}{n} - \left(\sum_{i=1}^{N}\sigma_i\right)^2\right\} = E_{M2}V_p(t_{lo})$$

The optimum estimator t_{lo} cannot be used in practice since in most situations, μ_i and $\sigma_i's$ are unknown for $i=1,...,N$.

Cassel, Särndal and Wretman (1977) also considered the model $M_{2:1}$ which is the model M2 with

$$\mu_i = a_i + \beta x_i \text{ and } \sigma_i^2 = \sigma^2 x_i^2$$

where a_i and x_i are positive and known $\left(\sum_{i=1}^{N} x_i = N\right)$ for $i=1,...,N$

but β, σ^2 and ρ are unknown

and $\frac{-1}{N-1} \le \rho \le 1$.

So under this model, the estimator t_{lo} reduces to

$$t_{lo:1} = N \sum_{i \in s} \frac{y_i - a_i}{n x_i} + a$$

where $a = \sum_{i=1}^{N} a_i$.

Thus we have the following optimal strategy using Theorem 5.4.6.

Corollary

Under the model $M_{2:1}$, $h_{lo} = (p_{\pi\sigma}, t_{lo})$ is optimal in the class of strategies $H = (p,t)$ in the sense that

$$E_{M_{2:1}} V_p(t) \geq (1-\rho) N^2 \left\{ 1 - \frac{\left(\sum_{i=1}^{N} x_i \right)^2}{N} \right\} \sigma^2 / n = E_{M_{2:1}} V_{p_{\pi\sigma}}(t_{lo:1}) \qquad p \in P_n,\ t \in C_{lu}$$

where $p_{\pi\sigma} (\in P_n)$ is a sampling design with inclusion probability of the ith unit $\pi_i = np_i, p_i = \frac{x_i}{X}$.

5.5 Optimal Model-unbiased estimators

A linear model unbiased (ξ-unbiased) estimator

$$t_s = a_s + \sum_{i \in s} b_{si} y_i \qquad (5.5.1)$$

for a finite population total Y satisfies $E_\xi(t_s) = E_\xi(Y)$.

The class of linear ξ-unbiased estimators will be denoted by $C_{\xi l}$.

The estimator t can be written as

$$t_s = \sum_{i \in s} y_i + a_s + \sum_{i \in s} (b_{si} - 1) y_i = \sum_{i \in s} y_i + t_s^*$$

where $t_s^* = a_s + \sum_{i \in s} (b_{si} - 1) y_i$ and $w_{si} = b_{si} - 1$.

The estimator t_s^* is a linear ξ-unbiased estimator for the unobserved total

$$Y\bar{s} = \sum_{i \notin s} y_i = \sum_{i \in \bar{s}} y_i$$

since $E_\xi(t_s^*) = E_\xi(Y\bar{s})$ where $\bar{s}$ consists of the units that do not belong to s.

Here we will find an optimal estimator t_s^0 in the class $C_{\xi l}$ for which $E_\xi E_p(t_s - Y)^2 = E_p E_\xi (t_s - Y)^2$ attains a minimum for a given design p under various superpopulation models.

Now $E_\xi E_p(t_s - Y)^2 = E_\xi E_p(t_s^* - Y\bar{s})^2$

$$= E_p[V_\xi(t_s^*) + V_\xi(Y\bar{s}) - 2C_\xi(t_s^*, Y\bar{s})]. \qquad (5.5.2)$$

Also if $C_\xi(y_i, y_j) = 0$ for $i \neq j$,

then $C_\xi(t_s^*, Y\bar{s}) = 0$

so that equation (5.4.2) becomes

$$E_\xi E_p(t_s - Y)^2 = E_p[V_\xi(t_s^*) + V_\xi(Y\bar{s})] = \sum_s [V_\xi(t_s^*) + V_\xi(\bar{Y}_s)]p(s). \qquad (5.5.3)$$

We thus conclude that for a given sampling design p, t_s becomes optimal by a suitable choice of b_{si} if for each s with $p(s) > 0$, $V_\xi(t_s^*)$ attains a minimum value among all linear ξ-unbiased estimators of $\bar{Y}_{\bar{s}}$.

The details of the model unbiased estimators are given by Cassel, Särndal and Wretman (1977).

5.6 Conclusion

We have seen in the earlier chapters that the design based approach often leads to no definite optimal strategy. To combat this problem, we have introduced the concept of superpopulation models in this chapter.

Inference under the model based approach allowed us to find best linear unbiased predictors. These predictors were then combined with suitable sampling designs to obtain an optimal strategy. This optimal strategy became a purposive sampling design.

We have also noted that a balanced sampling design should be used to ensure that we choose an appropriate model as model misspecification leads to inefficient estimators.

For design-unbiased estimators, we have shown the existence of an optimal estimator and have also presented several optimal strategies under two models. We have also seen that the Horvitz-Thompson (1952) estimator based on an appropriate sampling design becomes an optimal sampling strategy for various superpopulation models.

Finally we presented an optimal model-unbiased estimator.

Chapter 6

Some Specific Sampling Strategies

A sampling strategy is a combination of an estimator t and a sampling design p. The population under consideration is composed of N units from which a sample of size n is selected. We will denote the value of the study variable (y) and the auxiliary variable (x) for the units y_i and x_i respectively. Here it is assumed that the $x_i's$ are true for every $i = 1,...,N$.

In this chapter we will consider strategies which are commonly used in practice. This includes the Hansen-Hurwitz (1943) estimator based on PPSWR sampling scheme, Horvitz-Thompson (1952) estimator based on an arbitrary sampling scheme, the Midzuno-Sen sampling scheme and the Rao-Hartley- Cochran sampling strategy. The expressions of the variance and unbiased estimator of the variance have been provided.

Inclusion probability proportional to size sampling designs proposed by Brewer (1963), Durbin (1967) and Goodman and Kish (1950) have been also been presented.

We also compare performances of Rao-Hartley-Cochran sampling strategy, Midzuno-Sen sampling scheme and the Horvitz-Thompson estimator under a superpopulation models. Some numerical examples are also provided.

6.1 Probability Proportional to size with Replacement (PPSWR) Sampling Scheme

The units are selected independently at each draw. The probability of selecting the ith unit at any draw is $p_i = x_i / X$ $\left(X = \sum_i x_i ,\ p_i > 0 ,\ \sum_{i=1}^{n} p_i = 1 \right)$, which is called the normed size measure for the ith unit i.e. $p_i(k) = p_i$. So, the probability of selection of an ordered sample $s = (u_{i_1}, u_{i_2}, ..., u_{i_n})$ = $p_{i_1} ... p_{i_k}$.

6.1.1 Estimation of the population total and its variance

Let $y_{(r)}$ be the value of the study variable y, $x_{(r)}$ the value of the auxiliary variable x and $p_{(r)} = x_{(r)} / X$ be the normed size measure for the unit that is selected at the rth draw, $r = 1, ..., n$.

If the rth draw produces the ith unit then

$$P\left\{ \frac{y_{(r)}}{p_{(r)}} = \frac{y_i}{p_i} \right\} = p_i \qquad r = 1, ..., n \text{ and } i = 1, ..., N$$

Theorem 6.1

The estimator

$$\hat{Y}_{hh} = \frac{1}{n} \sum_{r=1}^{n} \frac{y_{(r)}}{p_{(r)}}$$

is known as the Hansen-Hurwitz estimator (Hansen-Hurwitz (1943)).

It follows that

i) $\hat{Y}_{hh}$ is an unbiased estimator of the population total $Y = \sum_{i=1}^{N} y_i$.

ii) The variance of $\hat{Y}_{hh}$ is $V\left(\hat{Y}_{hh}\right) = V_{PPS}/n$

where $V_{PPS} = \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y\right)^2 = \frac{1}{2}\sum_{i\neq}^{N}\sum_{j=1}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2$.

iii) An unbiased estimator of $V\left(\hat{Y}_{hh}\right)$ is

$$\hat{V}(\hat{Y}_{hh}) = \frac{1}{n(n-1)}\sum_{r=1}^{n}\left(\frac{y_{(r)}}{p_{(r)}} - \hat{Y}_{hh}\right)^2 .$$

Proof:

The expectation and variance of $\frac{y_{(r)}}{p_{(r)}}$ are computed as follows:

$$E\left(\frac{y_{(r)}}{p_{(r)}}\right) = \sum_{i=1}^{N}\frac{y_i}{p_i} p_i = Y \qquad \text{and} \tag{6.1.1}$$

$$\begin{aligned} V\left(\frac{y_{(r)}}{p_{(r)}}\right) &= E\left(\frac{y_{(r)}}{p_{(r)}} - Y\right)^2 \\ &= \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y\right)^2 \qquad (6.1.2) \\ &= \frac{1}{2}\sum_{i\neq}^{N}\sum_{j=1}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 \\ &= \frac{1}{2}\sum_{i\neq}^{N}\sum_{j=1}^{N} p_i \frac{y_i^2}{p_i} - \sum_{i\neq}^{N}\sum_{j=1}^{N} y_i y_j \\ &= \sum_{i=1}^{N} p_i \frac{y_i^2}{p_i}(1-p_i) - \left(Y^2 - \sum_{i=1}^{N} y_i^2\right) \qquad (6.1.3) \\ &= V_{PPS} . \end{aligned}$$

So we get:

i) $E(\hat{Y}_{hh}) = \frac{1}{n}\sum_{ri=1}^{n} E\left(\frac{y_{(r)}}{p_{(r)}}\right) = Y$.

ii) $V(\hat{Y}_{hh}) = \frac{1}{n^2} V\left(\sum_{r=1}^{n} \frac{y_{(r)}}{p_{(r)}}\right)$

$$= \frac{1}{n^2}\left\{V\left(\sum_{r=1}^{n} \frac{y_{(r)}}{p_{(r)}}\right) + \sum_{r\neq}^{n}\sum_{t}^{n} Cov\left(\frac{y_{(r)}}{p_{(r)}}, \frac{y_{(t)}}{p_{(t)}}\right)\right\}$$

$$= V_{PPS}/n$$

since $V\left(\frac{y_{(r)}}{p_{(r)}}\right) = V_{PPS}$ and $Cov\left(\frac{y_{(r)}}{p_{(r)}}, \frac{y_{(t)}}{p_{(t)}}\right) = 0$ for $r \neq t$ as the draws are independent.

iii) $E\left[\hat{V}\left(\hat{Y}_{hh}\right)\right] = \frac{1}{n(n-1)} E\left[\sum_{r=1}^{n}\left(\frac{y_{(r)}}{p_{(r)}}\right)^2 - n(\hat{Y}_{hh})^2\right]$

$$= \frac{1}{n(n-1)}\left[\sum_{r=1}^{n} E\left(\frac{y_{(r)}}{p_{(r)}}\right)^2 - nE(\hat{Y}_{hh})^2\right]$$

$$= \frac{1}{n(n-1)}\left[n(V_{PPS} + Y^2) - n\{V(\hat{Y}_{hh}) + Y^2\}\right]$$

$$= \frac{1}{n(n-1)}\left[nV_{PPS} - nV(\hat{Y}_{hh})\right] = \frac{1}{n(n-1)}\left[n^2 V(\hat{Y}_{hh}) - nV(\hat{Y}_{hh})\right]$$

$$= V(\hat{Y}_{hh}) .$$

6.2 Horvitz-Thompson (1952) Estimator based on an Arbitrary Sampling Scheme

The estimator of Horvitz-Thompson (1952) (t_{HTE}) is defined in section 3.1.3.3 as follows:

$$t_{HTE} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_i \frac{y_i}{\pi_i} I_{si} .$$

Using Theorem 3.1, 3.2 and 3.3 from Chapter 3, we find

i) $E(t_{HTE}) = Y$ and

ii) $$V(t_{HTE}) = \sum_i y_i^2 \left(\frac{1}{\pi_i} - 1 \right) + \sum_{i \neq} \sum_j y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right).$$

For a fixed effective size sampling design

iii) $$V(t_{HTE}) = \frac{1}{2} \sum_{i \neq} \sum_j (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = V(t_{YG}).$$

6.2.1 An unbiased estimator for the variance $V(t_{HTE})$

An unbiased estimator of $V(t_{HTE})$ was proposed by Horvitz and Thompson (1952)

$$\hat{V}_1(t_{HTE}) = \sum_i \frac{y_i^2}{\pi_i} \left(\frac{1}{\pi_i} - 1 \right) + \sum_{i \neq} \sum_j \frac{y_i y_j}{\pi_{ij}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right).$$

An unbiased estimator for $V(t_{HTE})$ is given by

$$\hat{V}(t_{HTE}) = \sum_{i<}\sum_{j} \frac{\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 = \hat{V}_{YG}\,.$$

This estimator is called the Yates-Grundy (1953) estimator.

Remark:

The unbiased estimator $\hat{V}(t_{HTE})$ can be used for any sampling design with $\pi_{ij} > 0$ for $i \neq j$. The demerit of this estimator is that $\hat{V}(t_{HTE})$ can take on negative values. No simple sufficient condition for the non-negativity of the estimator $\hat{V}(t_{HTE})$ is known.

The estimator $\hat{V}_{YG}$ can be used only for a fixed effective size sampling design with $\pi_{ij} > 0$ for $i \neq j$. Sufficient conditions of non-negativity of the estimator $\hat{V}_{YG}$ is $\pi_i \pi_j > \pi_{ij}$ for $i \neq j$. Various sampling designs are available for which $\hat{V}_{YG}$ is found to be non-negative.

6.3 Midzuno-Sen Sampling Scheme

(Midzuno (1952), Sen (1953))

In this sampling scheme at the first draw, the ith unit is selected with probability p_i then the remaining $n-1$ units are selected by SRSWOR method from the $N-1$ units which were not chosen in the first draw.

The probability of selecting an unordered sample $\tilde{s}$ = ($u_{i_1},...,u_{i_n}$) is

$$p(\tilde{s}) = \sum_{i\in\tilde{s}} p_i \frac{1}{\binom{N-1}{n-1}} = \frac{x_{\tilde{s}}}{X}\frac{1}{M_1}$$

where $x_{\tilde{s}} = \sum_{r\in\tilde{s}} x_r$, $X = \sum_{i=1}^{N} x_i$, $M_1 = \sum_{\tilde{s}} I_{si}$ = $\binom{N-1}{n-1}$

$$I_{si} = \begin{cases} 1 & if \;\; i \in \tilde{s} \\ 0 & if \;\; i \notin \tilde{s}. \end{cases}$$

Theorem 6.3

Let $t_{MS} = \frac{y_{\tilde{s}}}{x_{\tilde{s}}} \cdot X$ with $y_{\tilde{s}} = \sum_{i\in\tilde{s}} y_i$.

Then

i) t_{MS} is an unbiased estimator of the population total Y ,

ii) $$V(t_{MS}) = \sum_{i=1}^{N} y_i^2(\tau_i - 1) + \sum_{i\neq}^{N}\sum_{j=1}^{N} y_i y_j(\tau_{ij} - 1) \qquad (6.3.1)$$

where $\tau_i = \frac{X}{M_1}\sum_{\tilde{s}} \frac{I_{si}}{x_{\tilde{s}}}$, $\tau_{ij} = \frac{X}{M_1}\sum_{\tilde{s}} \frac{I_{si}I_{sj}}{x_{\tilde{s}}}$ and

iii) an unbiased estimator for $V(t_{MS})$ is

$$\hat{V}(t_{MS}) = \sum_{i=1}^{N} I_{si} y_i^2(\tau_i - 1)/\pi_i + \sum_{i\neq}^{N}\sum_{j=1}^{N} I_{si}I_{sj} y_i y_j(\tau_{ij} - 1)/\pi_{ij} .$$

Proof:

i) $$E(t_{MS}) = E\left(\frac{y_{\tilde{s}}}{x_{\tilde{s}}}.X\right)$$

$$= X\sum_{s\in\varphi}\frac{y_{\tilde{s}}}{x_{\tilde{s}}}p(\tilde{s})$$

$$= \frac{1}{M_1}\sum_{\tilde{s}} y_{\tilde{s}}$$

$$= \frac{1}{M_1}\sum_{\tilde{s}}\sum_{i=1}^{N} I_{\tilde{s}i}\, y_i$$

$$= \frac{1}{M_1}\sum_{i=1}^{N} y_i \sum_{\tilde{s}} I_{\tilde{s}i}$$

$$= \sum_i y_i = Y\,.$$

ii) $$E(t_{MS})^2 - Y^2 = E\left(\frac{1}{M_1}\frac{1}{p(\tilde{s})}\sum_{i=1}^{N} I_{\tilde{s}i}y_i\right)^2 - Y^2$$

$$= \frac{1}{M_1^2}E\left(\frac{1}{p(\tilde{s})^2}\left\{\sum_i I_{si}{}^2 y_i^2 + \sum_{i\neq}\sum_j I_{si}I_{sj}y_iy_j\right\}\right) - Y^2$$

$$= \frac{1}{M_1^2}\sum_s \frac{1}{p(\tilde{s})}\left\{\sum_i I_{si}y_i^2 + \sum_{i\neq}\sum_j I_{si}I_{sj}y_iy_j\right\} - Y^2$$

$$= \frac{1}{M_1^2}\left\{\sum_i y_i^2\sum_{\tilde{s}}\frac{I_{\tilde{s}i}}{p(\tilde{s})} + \sum_{i\neq}\sum_j y_iy_j\sum_{\tilde{s}}\frac{I_{\tilde{s}i}I_{\tilde{s}j}}{p(\tilde{s})}\right\} - Y^2$$

$$= \sum_i y_i^2\left\{\frac{1}{(M_1)^2}\sum_{\tilde{s}}\frac{I_{\tilde{s}i}}{p(\tilde{s})} - 1\right\} + \sum_{i\neq}^{N}\sum_j^{N} y_iy_j\left\{\frac{1}{(M_1)^2}\sum_{\tilde{s}}\frac{I_{\tilde{s}i}I_{\tilde{s}j}}{p(\tilde{s})} - 1\right\}$$

$$= \sum_{i=1}^{N} y_i^2(\tau_i - 1) + \sum_{i\neq}^{N}\sum_{j=1}^{N} y_iy_j(\tau_{ij} - 1)$$

since $$\frac{1}{M_1^2}\sum_s\frac{I_{\tilde{s}i}}{p(\tilde{s})} = \frac{XM_1}{M_1^2}\sum_{\tilde{s}}\frac{I_{\tilde{s}i}}{x_s} = \frac{X}{M_1}\sum_{\tilde{s}}\frac{I_{\tilde{s}i}}{x_{\tilde{s}}}$$

and $$\frac{1}{M_1^2}\sum_s\frac{I_{\tilde{s}i}I_{\tilde{s}j}}{p(\tilde{s})} = \frac{XM_1}{M_1^2}\sum_{\tilde{s}}\frac{I_{\tilde{s}i}I_{\tilde{s}j}}{x_s} = \frac{X}{M_1}\sum_{\tilde{s}}\frac{I_{\tilde{s}i}I_{\tilde{s}j}}{x_{\tilde{s}}}\,.$$

iii) $$E\left[\hat{V}(t_{MS})\right] = E\left[\sum_{i=1}^{N} I_{si} y_i^2 (\tau_i - 1)/\pi_i + \sum_{i\neq}^{N}\sum_{j=1}^{N} I_{si} I_{sj} y_i y_j (\tau_{ij} - 1)/\pi_{ij}\right]$$

$$= V(t_{MS}),$$

which follows since $E(I_{si}) = \pi_i$ and $E(I_{si} I_{sj}) = \pi_{ij}$.

Note:

The estimator $\hat{V}(t_{MS})$ can take on negative values. Sufficient conditions for non-negativity of $\hat{V}(t_{MS})$ were proposed by Hanurav (1966), Rao (1967), and Chaudhuri & Arnab (1979). The details of the sufficient conditions are in a complex form.

Example 6.1

Consider the following data (Cochran (1977), p35) relating to family income (y) and family size (x) for N=6 families.

Table 6.1: Family income and size for 6 families

Family	1	2	3	4	5	6
Income(y)	62	62	87	65	58	92
Size(x)	2	3	3	5	4	7
Cum Total	2	5	8	13	17	24

We can select a sample of n=3 families using the Midzuno-Sen sampling scheme as follows:

The first unit is chosen using probability proportional to size sampling. We select a random number from a random number table (Cochran (1977), p19). The random number is 17, so the first unit chosen is 5.

We then have to select the remaining 2 units by the SRSWOR method from the 9 units 1, 2, 3, 4, 5, 7, 8, 9, 10 that were not selected in the first draw.

Using the random number table once again, the selected units are unit 5 and unit 1.

So the selected sample is s = { 6, 5, 1 }.

$$t_{MS} = \frac{y_{\tilde{s}}}{x_{\tilde{s}}} \times X$$

$$= \frac{92+58+62}{7+4+2} \times 24$$

= 391.38 and

$$V(t_{MS}) = \sum_{i=1}^{N} y_i^2 (\tau_i - 1) + \sum_{i \neq}^{N} \sum_{j=1}^{N} y_i y_j (\tau_{ij} - 1).$$

Here $M_1 = \sum_{\tilde{s}} I_{si} = \binom{N-1}{n-1} = \binom{6-1}{3-1} = 10$

and the total of X = 24

so $\frac{X}{M_1}$ = 2.4.

Now $\tau_i = \frac{X}{M_1} \sum_{\tilde{s}} \frac{I_{si}}{x_{\tilde{s}}}$, we thus obtain:

$$\tau_1 = 2.4\left(\frac{1}{x_1+x_2+x_3} + \frac{1}{x_1+x_2+x_4} + \frac{1}{x_1+x_2+x_5} + \frac{1}{x_1+x_2+x_6} + \frac{1}{x_1+x_3+x_4} + \frac{1}{x_1+x_3+x_5} \right.$$

$$\left. + \frac{1}{x_1+x_3+x_6} + \frac{1}{x_1+x_4+x_5} + \frac{1}{x_1+x_4+x_6} + \frac{1}{x_1+x_5+x_6} \right)$$

= 2.287559,

$$\tau_2 = 2.4\left(\frac{1}{x_2+x_1+x_3}+\frac{1}{x_2+x_1+x_4}+\frac{1}{x_2+x_1+x_5}+\frac{1}{x_2+x_1+x_6}+\frac{1}{x_2+x_3+x_4}+\frac{1}{x_2+x_3+x_5}\right.$$

$$\left.+\frac{1}{x_2+x_3+x_6}+\frac{1}{x_2+x_4+x_5}+\frac{1}{x_2+x_4+x_6}+\frac{1}{x_2+x_5+x_6}\right)$$

= 2.180892,

$$\tau_3 = 2.4\left(\frac{1}{x_3+x_1+x_2}+\frac{1}{x_3+x_1+x_4}+\frac{1}{x_3+x_1+x_5}+\frac{1}{x_3+x_1+x_6}+\frac{1}{x_3+x_2+x_4}+\frac{1}{x_3+x_2+x_5}\right.$$

$$\left.+\frac{1}{x_3+x_2+x_6}+\frac{1}{x_3+x_4+x_5}+\frac{1}{x_3+x_4+x_6}+\frac{1}{x_3+x_5+x_6}\right)$$

= 2.180892,

$$\tau_4 = 2.4\left(\frac{1}{x_4+x_1+x_2}+\frac{1}{x_4+x_1+x_3}+\frac{1}{x_4+x_1+x_5}+\frac{1}{x_4+x_1+x_6}+\frac{1}{x_4+x_2+x_3}+\frac{1}{x_4+x_2+x_5}\right.$$

$$\left.+\frac{1}{x_4+x_2+x_6}+\frac{1}{x_4+x_3+x_5}+\frac{1}{x_4+x_3+x_6}+\frac{1}{x_4+x_5+x_6}\right)$$

= 1.957792,

$$\tau_5 = 2.4\left(\frac{1}{x_5+x_1+x_2}+\frac{1}{x_5+x_1+x_3}+\frac{1}{x_5+x_1+x_4}+\frac{1}{x_5+x_1+x_6}+\frac{1}{x_5+x_2+x_3}+\frac{1}{x_5+x_2+x_4}\right.$$

$$\left.+\frac{1}{x_5+x_2+x_6}+\frac{1}{x_5+x_3+x_4}+\frac{1}{x_5+x_3+x_6}+\frac{1}{x_5+x_4+x_6}\right)$$

= 2.068988 and

$$\tau_6 = 2.4\left(\frac{1}{x_6+x_1+x_2}+\frac{1}{x_6+x_1+x_3}+\frac{1}{x_6+x_1+x_4}+\frac{1}{x_6+x_1+x_5}+\frac{1}{x_6+x_2+x_3}+\frac{1}{x_6+x_2+x_4}\right.$$

$$\left.+\frac{1}{x_6+x_2+x_5}+\frac{1}{x_6+x_3+x_4}+\frac{1}{x_6+x_3+x_5}+\frac{1}{x_6+x_4+x_5}\right)$$

= 1.753516.

Now

$\tau_{ij} = \frac{X}{M_1} \sum_{\tilde{s}} \frac{I_{si} I_{sj}}{x_{\tilde{s}}}$ so

$$\tau_{12} = 2.4\left[\frac{1}{x_1+x_2+x_3}+\frac{1}{x_1+x_2+x_4}+\frac{1}{x_1+x_2+x_5}+\frac{1}{x_1+x_2+x_6}\right] = 1.006667,$$

$$\tau_{13} = 2.4\left[\frac{1}{x_1+x_3+x_2}+\frac{1}{x_1+x_3+x_4}+\frac{1}{x_1+x_3+x_5}+\frac{1}{x_1+x_3+x_6}\right] = 1.006667,$$

$$\tau_{14} = 2.4\left[\frac{1}{x_1+x_4+x_2}+\frac{1}{x_1+x_4+x_3}+\frac{1}{x_1+x_4+x_5}+\frac{1}{x_1+x_4+x_6}\right] = 0.86961,$$

$$\tau_{15} = 2.4\left[\frac{1}{x_1+x_5+x_2}+\frac{1}{x_1+x_5+x_3}+\frac{1}{x_1+x_5+x_4}+\frac{1}{x_1+x_5+x_6}\right] = 0.936131,$$

$$\tau_{16} = 2.4\left[\frac{1}{x_1+x_6+x_2}+\frac{1}{x_1+x_6+x_3}+\frac{1}{x_1+x_6+x_4}+\frac{1}{x_1+x_6+x_5}\right] = 0.756044,$$

$$\tau_{21} = 2.4\left[\frac{1}{x_2+x_1+x_3}+\frac{1}{x_2+x_1+x_4}+\frac{1}{x_2+x_1+x_5}+\frac{1}{x_2+x_1+x_6}\right] = 1.006667,$$

$$\tau_{23} = 2.4\left[\frac{1}{x_2+x_3+x_1}+\frac{1}{x_2+x_3+x_4}+\frac{1}{x_2+x_3+x_5}+\frac{1}{x_2+x_3+x_6}\right] = 0.942797,$$

$$\tau_{24} = 2.4\left[\frac{1}{x_2+x_4+x_1}+\frac{1}{x_2+x_4+x_3}+\frac{1}{x_2+x_4+x_5}+\frac{1}{x_2+x_4+x_6}\right] = 0.818182,$$

$$\tau_{25} = 2.4\left[\frac{1}{x_2+x_5+x_1}+\frac{1}{x_2+x_5+x_3}+\frac{1}{x_2+x_5+x_4}+\frac{1}{x_2+x_5+x_6}\right] = 0.878095,$$

$$\tau_{26} = 2.4\left[\frac{1}{x_2+x_6+x_1}+\frac{1}{x_2+x_6+x_3}+\frac{1}{x_2+x_6+x_4}+\frac{1}{x_2+x_6+x_5}\right] = 0.716044,$$

$$\tau_{31} = 2.4\left[\frac{1}{x_3+x_1+x_2}+\frac{1}{x_3+x_1+x_4}+\frac{1}{x_3+x_1+x_5}+\frac{1}{x_3+x_1+x_6}\right] = 1.006667,$$

$$\tau_{32} = 2.4\left[\frac{1}{x_3+x_2+x_1}+\frac{1}{x_3+x_2+x_4}+\frac{1}{x_3+x_2+x_5}+\frac{1}{x_3+x_2+x_6}\right] = 0.942797,$$

$$\tau_{34} = 2.4\left[\frac{1}{x_3+x_4+x_1}+\frac{1}{x_3+x_4+x_2}+\frac{1}{x_3+x_4+x_5}+\frac{1}{x_3+x_4+x_6}\right] = 0.818182,$$

$$\tau_{35} = 2.4\left[\frac{1}{x_3+x_5+x_1}+\frac{1}{x_3+x_5+x_2}+\frac{1}{x_3+x_5+x_4}+\frac{1}{x_3+x_5+x_6}\right] = 0.878095,$$

$$\tau_{36} = 2.4\left[\frac{1}{x_3+x_6+x_1}+\frac{1}{x_3+x_6+x_2}+\frac{1}{x_3+x_6+x_4}+\frac{1}{x_3+x_6+x_5}\right] = 0.716044,$$

$$\tau_{41} = 2.4\left[\frac{1}{x_4+x_1+x_2}+\frac{1}{x_4+x_1+x_3}+\frac{1}{x_4+x_1+x_5}+\frac{1}{x_4+x_1+x_6}\right] = 0.86961,$$

$$\tau_{42} = 2.4\left[\frac{1}{x_4+x_2+x_1}+\frac{1}{x_4+x_2+x_3}+\frac{1}{x_4+x_2+x_5}+\frac{1}{x_4+x_2+x_6}\right] = 0.818182,$$

$$\tau_{43} = 2.4\left[\frac{1}{x_4+x_3+x_1}+\frac{1}{x_4+x_3+x_2}+\frac{1}{x_4+x_3+x_5}+\frac{1}{x_4+x_3+x_6}\right] = 0.818182,$$

$$\tau_{45} = 2.4\left[\frac{1}{x_4+x_5+x_1}+\frac{1}{x_4+x_5+x_2}+\frac{1}{x_4+x_5+x_3}+\frac{1}{x_4+x_5+x_6}\right] = 0.768182,$$

$$\tau_{46} = 2.4\left[\frac{1}{x_4+x_6+x_1}+\frac{1}{x_4+x_6+x_2}+\frac{1}{x_4+x_6+x_3}+\frac{1}{x_4+x_6+x_5}\right] = 0.641429,$$

$$\tau_{51} = 2.4\left[\frac{1}{x_5+x_1+x_2}+\frac{1}{x_5+x_1+x_3}+\frac{1}{x_5+x_1+x_4}+\frac{1}{x_5+x_1+x_6}\right] = 0.936131,$$

$$\tau_{52} = 2.4\left[\frac{1}{x_5+x_2+x_1}+\frac{1}{x_5+x_2+x_3}+\frac{1}{x_5+x_2+x_4}+\frac{1}{x_5+x_2+x_6}\right] = 0.878095,$$

$$\tau_{53} = 2.4\left[\frac{1}{x_5+x_3+x_1}+\frac{1}{x_5+x_3+x_2}+\frac{1}{x_5+x_3+x_4}+\frac{1}{x_5+x_3+x_6}\right] = 0.878095,$$

$$\tau_{54} = 2.4\left[\frac{1}{x_5+x_4+x_1}+\frac{1}{x_5+x_4+x_2}+\frac{1}{x_5+x_4+x_3}+\frac{1}{x_5+x_4+x_6}\right] = 0.768182,$$

$$\tau_{56} = 2.4\left[\frac{1}{x_5 + x_6 + x_1} + \frac{1}{x_5 + x_6 + x_2} + \frac{1}{x_5 + x_6 + x_3} + \frac{1}{x_5 + x_6 + x_4}\right] = 0.677473,$$

$$\tau_{61} = 2.4\left[\frac{1}{x_6 + x_1 + x_2} + \frac{1}{x_6 + x_1 + x_3} + \frac{1}{x_6 + x_1 + x_4} + \frac{1}{x_6 + x_1 + x_5}\right] = 0.756044,$$

$$\tau_{62} = 2.4\left[\frac{1}{x_6 + x_2 + x_1} + \frac{1}{x_6 + x_2 + x_3} + \frac{1}{x_6 + x_2 + x_4} + \frac{1}{x_6 + x_2 + x_5}\right] = 0.716044,$$

$$\tau_{63} = 2.4\left[\frac{1}{x_6 + x_3 + x_1} + \frac{1}{x_6 + x_3 + x_2} + \frac{1}{x_6 + x_3 + x_4} + \frac{1}{x_6 + x_3 + x_5}\right] = 0.716044,$$

$$\tau_{64} = 2.4\left[\frac{1}{x_6 + x_4 + x_1} + \frac{1}{x_6 + x_4 + x_2} + \frac{1}{x_6 + x_4 + x_3} + \frac{1}{x_6 + x_4 + x_5}\right] = 0.641429 \text{ and}$$

$$\tau_{65} = 2.4\left[\frac{1}{x_6 + x_5 + x_1} + \frac{1}{x_6 + x_5 + x_2} + \frac{1}{x_6 + x_5 + x_3} + \frac{1}{x_6 + x_5 + x_4}\right] = 0.677473.$$

So $V(t_{MS}) = \sum_{i=1}^{N} y_i^2(\tau_i - 1) + \sum_{i\neq}^{N}\sum_{j=1}^{N} y_i y_j(\tau_{ij} - 1)$

$$= \left[\left(62^2 \times (2.28756 - 1)\right) + \ldots + \left(92^2 \times (1.75352 - 1)\right)\right]$$

$$+ \left[\left(62 \times 62 \times (1.00667 - 1) + \ldots + \left(58 \times 92 \times (0.67747 - 1)\right)\right]$$

$$= 32447.41 + (-27679.5) = 4767.91.$$

6.4 Rao-Hartley-Cochran (1962) Sampling Scheme

In this sampling scheme, the population is first divided at random into n disjoint groups so that the number of units belonging to the jth group G_j is N_j ($j = 1,\ldots,n$), a pre-assigned number with $N = \sum_{j=1}^{N} N_j$.

One unit is then selected from each of the groups with probability proportional to its measure of size.

So if the unit u_{i_j} belongs to the jth group G_j, it is selected with probability

$$q_{i_j} = \frac{x_{i_j}}{\sum_{k \in G_j} x_k} = \frac{p_{i_j}}{P_j}$$

where $p_i = {}^{x_i}\!/_{X}$ and $P_j = \sum_{k \in G} p_k$ = the sum of the $p_k's$ for the group G_j.

If the units $u_{i_1},...,u_{i_j},...,u_{i_n}$ are selected from the groups $G_1,...,G_j,...,G_n$ respectively, then an estimator based on the above sampling scheme is given by

$$t_{RHC} = \sum_{j=1}^{n} \frac{y_{i_j}}{p_{i_j}} P_j .$$

Theorem 6.4.1

i) t_{RHC} is an unbiased estimator for the population total Y,

ii) $V(t_{RHC}) = \frac{\sum_{j=1}^{n} N_j^2 - N}{N(N-1)} \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y \right)^2$ and

iii) an unbiased estimator for $V(t_{RHC})$ is

$$\hat{V}(t_{RHC}) = \frac{\sum_{j=1}^{n} N_j^2 - N}{N^2 - \sum_{j=1}^{n} N_j^2} \sum_{i=1}^{n} P_i \left(\frac{y_{i_j}}{p_{i_j}} - t_{RHC} \right)^2 .$$

Proof:

Let $G = G_1,...,G_j,...,G_n$ and E_G, V_G, $E(./G)$ and $V(./G)$ denote the unconditional expectation over G, unconditional variance over G, conditional expectation for a given G and conditional variance for a given G respectively.

i) $$E(t_{RHC}) = E_G\left\{\sum_{j=1}^{n} E\left(\frac{y_{i_j}}{p_{i_j}} P_j / G\right)\right\} = E_G \sum_{j=1}^{n} Y_j = Y$$

where $Y_j = \sum_{k \in G_k} y_k$.

ii) $$V(t_{RHC}) = E_G\left[V\left(\sum_{j=1}^{n} \frac{y_{i_j}}{p_{i_j}} P_j / G\right)\right] + V_G\left\{E\sum_{j=1}^{n} \frac{y_{i_j}}{p_{i_j}} P_j / G\right\}. \qquad (6.4.1)$$

Now

$$V_G\left[E\sum_{j=1}^{n} \frac{y_{i_j}}{p_{i_j}} P_j / G\right] = V_G(Y) = 0 \qquad (6.4.2)$$

and

$$E_G\left(V\sum_{j=1}^{n} \frac{y_{i_j}}{p_{i_j}} P_j / G\right)$$

$$= E_G\left[\sum_{j=1}^{n} V\left(\frac{y_{i_j}}{p_{i_j}} P_j / G\right) + \sum_{j \neq}^{n}\sum_{k}^{n} Cov\left(\frac{y_{i_j}}{p_{i_j}} P_i, \frac{y_{i_k}}{p_{i_k}} / G\right)\right]$$

$$= E_G \sum_{j=1}^{n} V\left(\frac{y_{ij}}{p_{ij}} p_i / G\right).$$

Since $Cov\left(\frac{y_{i_j}}{p_{i_j}}P_i, \frac{y_{i_k}}{p_{i_k}} / G\right)$ = 0 as samples are selected independently from each other.

$$= \sum_{j=1}^{n} E_G\left[\sum_{k\neq j}\sum_{t\in G_j} p_t \frac{y_k^2}{p_k} - y_k y_t\right] \qquad (6.4.3)$$

Now noting that G_j is a random sample of size N_j selected from the population of N units by SRSWOR, we obtain

$$\sum_{j=1}^{n} E_G\left[\sum_{k\neq j}\sum_{t\in G_j}\left(p_t \frac{y_k^2}{p_k} - y_k y_t\right)\right]$$

$$= \sum_{j} \frac{N_j(N_j-1)}{N(N-1)} \sum_{i\neq}^{N}\sum_{k=1}^{N}\left(p_t \frac{y_k^2}{p_k} - y_k y_t\right)$$

$$= \sum_{j} \frac{N_j(N_j-1)}{N(N-1)} \sum_{i=1}^{N} p_i\left(\frac{y_i}{p_i} - Y\right)^2 . \qquad (6.4.4)$$

Substituting (6.4.2) and (6.4.4) in to (6.4.1) we prove part (ii).

iii) $E\sum_{j=1}^{N} P_j\left(\frac{y_{i_j}}{p_{i_j}} - t_{RHC}\right)^2$

$$= E\sum_{j=1}^{n} P_j \frac{y_i^2}{p_{i_j}^2} - E(t_{RHC}^2)$$

$$= E\sum_{j=1}^{n} P_j \frac{y_i^2}{p_{i_j}^2} - V(t_{RHC}) - Y^2 . \qquad (6.4.5)$$

Now $$E\sum_{j=1}^{n} P_j \frac{y_i^2}{p_{i_j}^2}$$

$$= E_G \sum_{j=1}^{n} E(\frac{y_i^2}{p_{i_j}^2}.P_j / G)$$

$$= \sum_{j=1}^{n} E_G (\sum_{i \in G_j} \frac{y_i^2}{p_i^2})$$

$$= \sum_{j=1}^{n} \frac{N_j}{N} \sum_{i=1}^{N} \frac{y_i^2}{p_i^2}$$

$$= \sum_{i=1}^{N} \frac{y_i^2}{p_i^2}. \qquad (6.4.6)$$

So substituting (6.4.6) into (6.4.5) we get

$$E\sum_{j=1}^{N} P_j \left(\frac{y_{i_j}}{p_{i_j}} - t_{RHC} \right)^2$$

$$= \sum_{i=1}^{N} P_i \left(\frac{y_i}{p_i} - Y \right)^2 - V(t_{RHC})$$

$$= \left[\frac{N(N-1)}{\sum_j N_j^2 - N} - 1 \right] V(t_{RHC}). \qquad (6.4.7)$$

So from (6.4.7), we get $E[\hat{V}(t_{RHC})] = V(t_{RHC})$.

By Cauchy's inequality

$$n\sum N_i^2 \geq \left(\sum N_i\right)^2 = N^2$$

hence $$\sum_{i=1}^{N} N_i^2 \geq \frac{N^2}{n}$$

and $\sum N_i^2$ minimum when $N_i = \frac{N}{n}$.

Theorem 6.4.2

Assuming $N_i = \frac{N}{n}$ is an integer, we get

$$\text{i)}\, V(t_{RHC}) = \frac{N-n}{n(N-1)} \sum p_i \left(\frac{y_i}{p_i} - Y \right)^2$$

and

$$\text{ii)}\, \hat{V}(t_{RHC}) = \frac{N-n}{N(n-1)} \sum_{i=1}^{n} p_i \left(\frac{y_i}{p_i} - t_{RHC} \right)^2 .$$

Proof:

The theorem can be proved by putting $N_j = \frac{N}{n}$ in the above theorem.

i) $$V(t_{RHC}) = \frac{\sum_{j=1}^{n} N_j^2 - N}{N(N-1)} \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y \right)^2$$

$$= \frac{\sum_{j=1}^{n} \left(\frac{N}{n} \right)^2 - N}{N(N-1)} \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y \right)^2$$

$$= \frac{\frac{N^2}{n} - N}{N(N-1)} \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y \right)^2$$

$$= \frac{N-n}{n(N-1)} \sum p_i \left(\frac{y_i}{p_i} - Y \right)^2$$

ii) $$\hat{V}(t_{RHC}) = \frac{\sum_{j=1}^{n} N_j^2 - N}{N^2 - \sum_{j=1}^{n} N_j^2} \sum_{i=1}^{N} P_i \left(\frac{y_{i_j}}{p_{i_j}} - t_{RHC} \right)^2$$

$$= \frac{\sum_{j=1}^{n} \left(\frac{N}{n} \right)^2 - N}{N^2 - \sum_{j=1}^{n} \left(\frac{N}{n} \right)^2} \sum_{i=1}^{N} P_i \left(\frac{y_{i_j}}{p_{i_j}} - t_{RHC} \right)^2$$

$$= \frac{\frac{N^2}{n} - N}{N^2 - \frac{N^2}{n}} \sum_{i=1}^{N} P_i \left(\frac{y_{i_j}}{p_{i_j}} - t_{RHC} \right)^2$$

$$= \frac{N-n}{N(n-1)} \sum_{i=1}^{n} p_i \left(\frac{y_i}{p_i} - t_{RHC} \right)^2$$

Remark:

i) The variance $\hat{V}(t_{RHC})$ is always non-negative.

ii) The Rao-Hartley-Cochran estimator t_{RHC} is inadmissible because it is based on the order of the selection of units.

iii) The Rao-Hartley-Cochran estimator t_{RHC} is more efficient than the Horvitz-Thomson estimator t_{HTE} because $V(t_{RHC}) < V(t_{HTE})$.

Example 6.2

Referring to Example 6.1, we have the following data relating to family income (in 1000's) and family. We want to select a sample of size n=3 from a population of size N=10 using the Rao-Hartley-Cochran sampling strategy.

Table 6.2a: Family income and size of 6 families

Family	1	2	3	4	5	6	7	8	9	10
Income(y)	62	62	87	65	58	92	88	79	83	62
Size(x)	2	3	3	5	4	7	2	4	2	5

The first step is to randomly divide the population into n =3 groups. Using the random number table (Cochran (1977), p19), we get the following groups

Table 6.2b: Families grouped into 3 groups

Group	G_1			G_2				G_3		
Family	3	4	6	1	5	7	9	2	8	10
Income(y)	87	65	92	62	58	88	83	62	79	62
Size(x)	3	5	7	2	4	2	2	3	4	5
Cum Total	3	8	15	2	6	8	10	3	7	12

We now select one unit from each of the groups with probability proportional to its measure of size.

Using 2 columns in the random number table (Cochran (1977), p19) we select the units as follows:

Random number	**Unit**
02	3
13	-
78	-
16	-
65	-
01	1
15	-
11	10

So the selected sample is s = { 1, 3, 10 }.

So $t_{RHC} = \sum_{j=1}^{n} \frac{y_{i_j}}{p_{i_j}} P_j$ where $p_i = x_i / X$ and $P_j = \sum_{k \in G} p_k$

$$= \left(\frac{62}{2/37} \cdot \frac{10}{37}\right) + \left(\frac{87}{3/37} \cdot \frac{15}{37}\right) + \left(\frac{62}{5/37} \cdot \frac{12}{37}\right)$$

$$= 893.8,$$

$$V(t_{RHC}) = \frac{N-n}{n(N-1)} \sum p_i \left(\frac{y_i}{p_i} - Y\right)^2$$

$$= \frac{10-3}{3(10-1)} \times$$

$$\left(\frac{3}{37}\left(\frac{87}{3/37} - 738\right)^2 + \frac{5}{37}\left(\frac{65}{5/37} - 738\right)^2 + .. + \frac{4}{37}\left(\frac{79}{4/37} - 738\right)^2 + \frac{5}{37}\left(\frac{62}{5/37} - 738\right)^2\right)$$

$$= 34024.14.$$

and

$$\hat{V}(t_{RHC}) = \frac{N-n}{N(n-1)} \sum_{i=1}^{n} P_i \left(\frac{y_i}{p_i} - t_{RHC}\right)^2$$

$$= \frac{10-3}{3(10-1)}\left(\frac{2}{37}\left(\frac{62}{2/37} - 893.8\right)^2 + \frac{3}{37}\left(\frac{87}{3/37} - 893.84\right)^2 + \frac{5}{37}\left(\frac{62}{5/37} - 893.84\right)^2\right)$$

$$= 23778.22.$$

6.5 Inclusion Probability Proportional to Measure of Size Sampling Scheme (IPPS or πps)

The Horvitz-Thompson estimator, t_{HTE}, based on a fixed sample size design becomes constant if the $y_i's$ are proportional to the inclusion probabilities $\pi_i's$ and in this case the variance becomes zero.

The values of the $y_i's$ are unknown before the survey so one cannot construct a sampling design with inclusion probabilities that are proportional to y_i values.

If an auxiliary variable with values that are positive, known and approximately proportional to the study variable y is available, the variance of t_{HTE} is expected to be small for a sampling design whose inclusion probability is proportional to the measure of size i.e.

$$\pi_i = n\frac{x_i}{X} = np_i \qquad \text{as } p_i = \frac{x_i}{X}.$$

A sampling design is said to be an IPPS or πps sampling design if

i) $\pi_i = np_i < 1$ i.e. $p_i < \frac{1}{n}$ for every $i \in U$

ii) $\pi_{ij} > 0$ for $i, j \in U$.

Several IPPS sampling schemes are available in literature, but most of them are very complex.

6.5.1 Brewer's (1963) Sampling Design (n=2)

In this method, the ith unit is selected at the first draw with probability

$$p_i(1) = \frac{2p_i(1-p_i)}{A(1-2p_i)}$$

where $$A = \sum_{i\in U}\frac{2p_i(1-p_i)}{(1-2p_i)} = \sum_{i\in U}\frac{p_i(1+1-2p_i)}{(1-2p_i)} = 1+\sum_{i\in U}\frac{p_i}{(1-2p_i)}. \qquad (6.5.1)$$

The conditional probability of selecting the jth unit in the second draw given that the ith unit is selected at the first draw is

$$p_{j|i}(2) = \frac{p_j}{(1-p_i)} \qquad \text{for } j \neq i \in U$$

and

$$p_{i|i}(2) = 0 .$$

The inclusion probability of the ith unit is

$$\begin{aligned}
\pi_i &= p_i(1) + \sum_{j \neq i} p_j(1)\, p_{i|j}(2) \\
&= \frac{2p_i}{A}\left\{\frac{1-p_i}{1-2p_i} + \sum_{j \neq 1} \frac{p_j}{1-2p_j}\right\} \\
&= 2p_i .
\end{aligned}$$

The inclusion probability for the ith and jth unit ($i \neq j$) is

$$\begin{aligned}
\pi_{ij} &= p_i(1)\, p_{j|i}(2) + p_j(1)\, p_{i|j}(2) \\
&= \frac{2p_i p_j}{A}\left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right).
\end{aligned}$$

So the difference is given by

$$\begin{aligned}
\pi_i \pi_j - \pi_{ij} &= \frac{2p_i p_j}{A}\left(2A - \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right)\right) \\
&= \frac{2p_i p_j}{A} \sum_{k \neq (i,j)} \frac{p_k}{1-2p_k} \\
&> 0 .
\end{aligned}$$

6.5.2 Durbin's (1967) Sampling Design (n=2)

In this sampling scheme, the probability of selecting the ith unit at the first draw is

$$p_i(1) = p_i \qquad \text{for } i \in U .$$

The conditional probability of selecting the jth unit given that the ith unit was selected at the first draw is

$$p_{j|i}(2) = p_j \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j} \right) \Big/ A \qquad \text{for } j \neq i$$

and

$$p_{i|i}(2) = 0$$

where A is given in (6.5.1) above and $\sum_{j \in U} p_{j|i} = 1.$

The probability of selecting an unordered sample (i, j) is

$$\begin{aligned} \pi_{ij} &= p_i(1)\, p_{j|i}(2) + p_j(1)\, p_{i|j}(2) \\ &= 2 p_i p_j \left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j} \right) \Big/ A \end{aligned}$$

which is the inclusion probability of the ith and jth unit for Brewer's (1963) sampling scheme.

The inclusion probability for the ith unit is

$$\pi_i = \sum_{j(\neq i)} \pi_{ij} = 2 p_i .$$

So the difference $\pi_i \pi_j - \pi_{ij} > 0$.

6.5.3 Goodman and Kish (1950)

In this sampling procedure, we assume $np_i \leq 1$ for every $i \in U$.

Let $\Gamma_i = n\sum_{j=1}^{n} p_j$ for $i = 1,...,n$

and $\Gamma_0 = 0$.

A random start d is selected from a uniform distribution over (0,1). The random start selects sample units whose index "j" satisfies

$$\Gamma_{j-1} \leq d + k < \Gamma_j \qquad \text{for } k = 0,...,n-1.$$

This sampling procedure can be used for the selection of an IPPS sample for any value of n.

No simple expression for π_{ij} is available. Hartley and Rao (1978) gave an expression for π_{ij} . An approximate expression for the variance of the Horvitz-Thompson estimator, $t_{HTE} = \sum_{i \in s} \frac{y_i}{\pi_i}$ is provided by Ashok and Sukhatme (1976).

$$\frac{1}{n}\left(\sum_i p_i z_i^2 - (n-1)\sum_i p_i^2 z_i^2\right) - \frac{n-1}{n}\left(2\sum_i p_i^3 z_i^2 - \sum_i p_i^2 \sum_i p_i^2 z_i^2 - 2(\sum_i p_i^2 z_i^3)^2\right)$$

$= V_{GK}$. (6.5.2)

Where $z_i = \frac{y_i}{p_i} - Y$ and V_{GK} is called the Goodman-Kish estimator.

An unbiased estimator of V_{GK} is

$$\hat{V}_{GK} = \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

$$= \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\left(\frac{n^2 p_i p_j}{\pi_{ij}} - 1\right)\left(\frac{y_i}{np_i} - \frac{y_j}{np_j}\right)^2 .$$

The expression (6.5.2) above indicates that the variance of t_{HTE} based on the Goodman and Kish sampling design provides a smaller variance than the Hansen-Hurwitz estimator based on PPSWR sampling.

6.6 Comparison of Strategies under Super Population Models

Here we compare the Horvitz-Thompson estimator with IPPS sampling design, the Rao-Hartley–Cochran strategy and the Midzuno-Sen strategy. These strategies are most commonly used in practice. This comparison is done using the following superpopulation model.

Superpopulation model M:

$$E_M(y_i) = \beta x_i ,$$

$$V_M(y_i) = \sigma^2 x_i^g \text{ and}$$

$$Cov_M(y_i, y_j) = 0 \qquad \text{for } i \neq j \qquad (6.6.1)$$

where $\beta, \sigma^2 (>0)$ are unknown constants, g is unknown but anticipated to lie in the interval (0, 2). Here the $x_i's$ are positive known constants.

E_M, V_M and Cov_M denote respectively, the expected value, variance and covariance with respect to the model M.

The model (6.6.1) was used by Cochran (1963), Cassel, Särndal and Wretman (1977), Rao (1967), Hanurav (1967), Chaudhuri & Arnab (1979) among others.

The variance of the Horvitz-Thompson estimator, $t_{HTE} = \sum_{i \in s} \frac{y_i}{\pi_i}$ is given by

$$V(t_{HTE}) = \sum_i y_i^2 \left(\frac{1}{\pi_i} - 1 \right) + \sum_{i \neq} \sum_j y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right).$$

The expected variance of the t_{HTE} is given by

$$\begin{aligned} E_M V(t_{HTE}) &= \sum E_M (y_i^2) \left(\frac{1}{\pi_i} - 1 \right) + \sum_{i \neq} \sum_j E_M (y_i y_j) \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) \\ &= \sum \left(\beta^2 x_i^2 + \sigma^2 x_i^g \right) \left(\frac{1}{\pi_i} - 1 \right) + \beta^2 \sum_{i \neq} \sum_j x_i x_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right). \end{aligned}$$

For a πps sampling scheme $\pi_i = np_i = n x_i / X$.

$$\begin{aligned} E_M V(t_{HTE}) &= \sum \sigma^2 x_i^g \left(\frac{1}{\pi_i} - 1 \right) \qquad (6.6.2) \\ &= E_1. \end{aligned}$$

The variance of the Rao-Hartley-Cochran estimator with N/n as an integer is given by

$$\begin{aligned} V(t_{RHC}) &= \frac{N-n}{n(N-1)} \sum p_i \left(\frac{y_i}{p_i} - Y \right)^2 \\ &= \frac{N-n}{n(N-1)} \left[\sum \frac{y_i^2}{p_i} - Y^2 \right]. \end{aligned}$$

And the expected variance of t_{RHC} is given by

$$EV(t_{RHC}) = \frac{N-n}{n(N-1)}\left[\sum \frac{\beta^2 x_i^2 + \sigma^2 x_i^g}{p_i} - \sigma^2 x_i^g - \beta^2 x_i^2\right].$$

Now putting $p_i = {x_i}/{X}$, we get

$$E_M V(t_{RHC}) = \frac{N-n}{n(N-1)}\left[\sigma^2\left(X\sum x_i^{g-1} - \sum x_i^g\right)\right] \qquad (6.6.3)$$

$$= E_2 .$$

The variance for the estimator t_{MS} based on the Midzuno-Sen sampling scheme is given by

$$V(t_{MS}) = \sum y_i^2(\tau_i - 1) + \sum\sum y_i y_j(\tau_{ij} - 1)$$

where $$\tau_i = \frac{X}{\binom{N-1}{n-1}}\sum_i \frac{1}{x_i + x_{i_2} + ... + x_{i_n}}$$

and $$\tau_{ij} = \frac{X}{\binom{N-1}{n-1}}\sum_{i,j} \frac{1}{x_i + x_j + x_{i_3} + ... + x_{i_n}}.$$

If $\sum_i$ and $\sum_{i,j}$ denote the summation over $n-1$ distinct numbers $(i_2,...,i_n)$ other than i and the summation over $n-2$ distinct numbers $(i_3,...,i_n)$ other than i and j respectively.

Also, it is known know that $\sum \tau_i X_i = \frac{N}{n}X$.

The expected variance for t_{MS} is given by

$$\begin{aligned} E_M V(t_{MS}) &= \beta^2\left[\sum x_i^2(\tau_i - 1) + \sum\sum x_i x_j(\tau_{ij} - 1)\right] + \sigma^2 \sum x_i^g(\tau_i - 1) \\ &= \sigma^2 \sum x_i^g(\tau_i - 1) \qquad (6.6.4) \\ &= E_3 . \end{aligned}$$

Since $\sum x_i^2(\tau_i - 1) + \sum\sum x_i x_j(\tau_{ij} - 1) = 0$.

6.6.1 Comparison between the Horvitz-Thompson Estimator and the Rao-Hartley-Cochran Strategy

Following Hanurav (1967), we get

$$\begin{aligned} E_2 - E_1 &= \frac{n-1}{n(N-1)}\sigma^2\left[N\sum_{i=1}^{N} x^g{}_i - X\sum_{i=1}^{N} x_i^{g-1}\right] \\ &= \frac{n-1}{n(N-1)}\sigma^2 N\left[\sum_{i=1}^{N} x_i^g\left(x_i - \frac{X}{n}\right)\right] \\ &= \frac{n-1}{n(N-1)}\sigma^2 N^2 Cov\left[x_i^{g-1}, x_i\right]. \end{aligned}$$

So	$E_2 - E_1 \geq 0$	if	$g - 1 > 0$ i.e. $g > 1$,	
	$E_2 - E_1 = 0$	if	$g - 1 = 0$ i.e. $g = 1$ and	
	$E_2 - E_1 \leq 0$	if	$g - 1 < 0$ i.e. $g < 1$.	(6.6.5)

Thus the Horvitz-Thompson estimator is superior to the Rao-Hartley-Cochran strategy under the superpopulation model M when $g > 1$. For $g < 1$, the Rao-Hartley-Cochran strategy is better than the Horvitz-Thompson estimator. The two strategies are equally efficient for $g = 1$.

6.6.2 Comparison between the Horvitz-Thompson estimator and the Midzuno-Sen strategy

Following Rao (1967), we get

$$E_3 - E_1 = \sigma^2 \sum_{i=1}^{N} x_i^{g-1} \left[\tau_i x_i - \frac{1}{N} \sum_{i=1}^{N} \tau_i x_i \right]$$
$$= \sigma^2 N Cov\left[\tau_i x_i, x_i^{g-1} \right].$$

Rao (1967) showed that $\tau_i x_i$ is an increasing function of x_i and x_i^{g-1} increases when $g > 1$ so in this case $E_3 - E_1 > 0$.

On the other hand for $g < 1$, x_i^{g-1} decreases as x_i increases but as x_i increases, $\tau_i x_i$ decreases.

Hence for $g < 1$ $E_3 - E_1 < 0$ and for $g = 1$, $x_i^{g-1} = 1$ so we have $E_3 - E_1 = 0$.i.e.

$$\begin{aligned} E_3 - E_1 &> 0 && \text{for } g > 1, \\ E_3 - E_1 &= 0 && \text{for } g = 1 \text{ and} \\ E_3 - E_1 &< 0 && \text{for } g < 1. \end{aligned} \qquad (6.6.6)$$

Thus the Horvitz-Thompson estimator is better than the Midzuno-Sen strategy for $g > 1$. For $g < 1$, the Midzuno-Sen strategy is better than the Horvitz-Thompson strategy. For $g = 1$ both strategies are equally efficient.

6.6.3 Comparison between the Midzuno-Sen Strategy and the Rao-Hartley-Cochran Strategy

Following Chaudhuri and Arnab (1979), we get

$$E_2 - E_3 = \sigma^2 \sum x_i^{g-1} z_i$$

where $z_i = \frac{N-n}{n(N-1)} X - \tau_i x_i + \frac{N-n}{n(N-1)} x_i$.

So that $\sum z_i = 0$.

Hence

$$E_2 - E_3 = \sigma^2 N Cov\left[x_i^{g-1}, z_i\right].$$

It follows that

$$\frac{\partial z_i}{\partial x_i} = 1 - \left[\tau_i + x_i \frac{\partial \tau_i}{\partial x_i}\right].$$

Now

$$\tau_i + x_i \frac{\partial \tau_i}{\partial x_i} =$$

$$\frac{1}{\binom{N-1}{n-1}}\left[X \sum_i \frac{1}{x_i + x_{i_2} + \ldots + x_{i_n}} + X_i \left\{\sum_i \frac{1}{x_i + x_{i_2} + \ldots + x_{i_n}} - X \sum_i \frac{1}{\left(x_i + x_{i_2} + \ldots + x_{i_n}\right)^2}\right\}\right]$$

$$= \frac{1}{\binom{N-1}{n-1}} \sum_i \frac{x_i\left(x_i + x_{i_2} + \ldots + x_{i_n}\right) + \left(x_{i_2} + \ldots + x_{i_n}\right)X}{\left(x_i + x_{i_2} + \ldots + x_{i_n}\right)^2}$$

$$> \frac{1}{\binom{N-1}{n-1}} \sum_i \frac{x_i\left(x_i + x_{i_2} + \ldots + x_{i_n}\right) + \left(x_{i_2} + \ldots + x_{i_n}\right)\left(x_i + x_{i_2} + \ldots + x_{i_n}\right)}{\left(x_i + x_{i_2} + \ldots + x_{i_n}\right)^2} = 1.$$

This implies that $\frac{\partial z_i}{\partial x_i} < 0$, i.e. z_i is a decreasing function of x_i.

So clearly

$$\begin{aligned} E_2 - E_3 &< 0 && \text{if } g > 1, \\ E_2 - E_3 &= 0 && \text{if } g = 1 \text{ and} \\ E_2 - E_3 &> 0 && \text{if } g < 1. \end{aligned} \qquad (6.6.7)$$

Thus the Midzuno-Sen strategy is better than the Rao-Hartley-Cochran strategy if $g > 1$. If $g < 1$, then the Rao-Hartley-Cochran strategy is more efficient. Both strategies are equally efficient if $g = 1$.

Now combining (6.7.5), (6.7.6) and (6.7.7), we get the following theorem.

Theorem 6.6.1

For the superpopulation model M

$$\begin{aligned} E_1 < E_2 < E_3 && \text{if } g > 1, \\ E_1 > E_2 > E_3 && \text{if } g < 1 \text{ and} \\ E_1 = E_2 = E_3 && \text{if } g = 1. \end{aligned}$$

6.7 Conclusion

The probability proportional to size with replacement sampling scheme (PPSWR) is easy to execute. The expressions of the Hansen-Hurwitz estimator, its variance and the unbiased estimator of its variance are very elegant and easy to compute. The main drawback of the Hansen-Hurwitz estimator based on PPSWR sampling is that it is inadmissible. Rao-

Blackwellization of the Hansen-Hurwitz estimator does not yield any elegant expression in general and hence cannot be used.

The Rao-Hartley-Cochran sampling scheme is also easy to execute the expression of its variance and unbiased estimator of variance are elegant. It is more efficient than the Hansen-Hurwitz estimator based on PPSWR sampling. The main drawback of Rao-Hartley-Cochran estimator it is that it is inadmissible. Rao-Blackwellization of the Rao-Hartley-Cochran estimator does not yield any elegant result.

The Midzuno-Sen sampling scheme is very easy to use, expressions of the unbiased estimator, variance and unbiased estimator of variance are easily available. The main drawback is that we may get non-negative variance estimates in all situations.

IPPS sampling scheme for a sample size n greater than 2 is in general very difficult to execute. The easiest is the Goodman Kish sampling procedure (section 6.5.3). The main demerit of this is the complexity of the expression of the second order inclusion probabilities.

The comparison between Rao-Hartley-Cochran, Horvitz-Thompson and Midzuno- Sen sampling strategies reveals that one should use the Horvitz Thompson estimator if $g > 1$ and the Rao-Hartley-Cochran estimator if $g < 1$. Obviously one needs to test the suitability of the model before using the estimator.

Chapter 7

Conclusion

The aim of this thesis was to present some inferential aspects when sampling from a finite population. The first step before any inference can be done is the selection of the sample. The methods of selection that were considered in this thesis were the cumulative total method, a sampling design and Hanurav's algorithm. Hanurav (1966) first established the relationship between a sampling scheme and a sampling design. His findings are very useful in the selection of a sample according to a sampling design.

After the selection of the sample we collect data $d = (y_i, i \in s)$ and make inference of the population parameter. Here y_i is the value of the character (y) under study for the ith $(i = 1,..,N)$ unit of the population. Our objective is to estimate some parametric function of the population. After collecting the data, we only know $y_i, i \in s$ but we do not know $y_i, i \notin s$. So in making inference from a finite population, we establish a link between $y_i, i \in s$ and $y_i, i \notin s$. There is no unique method to establish a link for a finite population. We normally use three methods. They are (i) design based approach, (ii) model based approach and (iii) model-design based approach.

In design based inference the link is established through a sampling design. Godambe (1955) established the non-existence theorem. Godambe's result was extended by Basu (1971).The unexpected non-existence theorem has tremendous implications for the inferential aspects of finite population sampling as for a given sampling design, we can construct infinitely many unbiased

estimators but we cannot choose any of them having the lowest variance in all the situations.

To eliminate inefficient estimators, the concept of admissible estimators has been introduced. Various admissible estimators exist for estimating a finite population total for a given sampling design. The concept of hyper admissibility was proposed by Hanurav (1965, 1968) to choose among other admissible estimators. However, some estimators are inadmissible. These estimators may be improved using the concept of sufficiency in finite population sampling and the "Rao-Blackwell" theorem.

In model based inference, the finite population vector $\underline{y} = (y_1,..,y_N)$ is assumed to be the realized outcome of a random variable $\underline{Y} = (Y_1,..,Y_N)$. The joint distribution of $\underline{Y}$ has been denoted by ξ. The unknown and unobserved values of the $y_i's$ is predicted by using the observed $d = (y_i, i \in s)$ through the superpopulation model ξ. In this model based approach an optimum estimator for some of the population parametric functions exist, this optimum estimator however is highly dependent on the model chosen. If an inappropriate model is chosen, the optimum estimator may not perform well. This problem may be overcome by using a balanced sampling design. However a balanced sampling design may not always be available.

The model-design based approach is a hybrid of the design based and the model based approach. In this approach inference is based on the assumed superpopulation model and sampling design. It is expected that model design based inference also protects against model misspecification. In this approach optimum sampling strategies for estimating finite population total under various superpopulation models exists.

We then consider a few sampling strategies that are commonly used in practice and provide expressions for estimators of the population total, the variance and an unbiased estimator of the variance. We also compare the performances of the Rao-Hartley-Cochran, Horvitz-Thompson and Midzuno-Sen sampling strategies.

Finally it should be noted that this thesis only discusses the theory of point estimation. The problem of interval estimation of the parametric functions such as the population mean, variance etc. was not discussed. The problem of optimum estimation of the sample size has also not been discussed. In interval estimation and optimum sample size determination, one is required to estimate the variance of the concerned estimator. The choice of an estimator with minimum variance is thus not enough. The variance of the chosen estimator should have additional properties such as (i) an elegant expression of variance, which can be used in practice; (ii) the unbiasedness property and (iii) the non-negativity property of the variance estimators. The nonnegative property is essential for the determination of a confidence interval as well as sample size.

So to sum up, this thesis has presented some inferential aspects when sampling from a finite population. The first thing that we looked at was the selection of a sample using the cumulative total method, a sampling design and Hanurav's algorithm. Once the sample is selected we wish to estimate a parametric function of interest. To do this we need to find a link between known observed data and unknown unobserved data. The following three methods were considered in this thesis:

i) **the design based approach** – here the link is established through a sampling design. A problem with this approach is the non-existence of an MVUE (Godambe (1955) and Basu (1971)). Admissibility of estimators can be used to eliminate inefficient estimators. However some estimators are inadmissible. These estimators may be improved using the concept of sufficiency and Rao-Blackwellisation.

ii) **the model based approach** – here a superpopulation model is used to predict unknown values. Many optimal estimators can be found but they are highly dependent on the model that was chosen so an incorrect model can lead to an inefficient estimator. Balanced sampling can be used to overcome this problem.

iii) **the model-design based approach**- inference is based on a superpopulation model and a sampling design. This type of inference protects against model misspecification. Many optimal strategies for estimating the finite population total exist.

Finally we looked at the estimation of the population total, the variance and an unbiased estimator of the variance for some specific sampling strategies. We also compared the efficiency of three commonly used strategies by calculating and comparing the expected variance of their estimators. The comparison between the Rao-Hartley-Cochran, Horvitz-Thompson and Midzuno- Sen sampling strategies reveals which estimator might be suitable for different values of g (equation 6.1).

Bibliography

1. Arnab, R. 2006. Survey Sampling Theory, *unpublished manuscript consulted by the courtesy of the author*.

2. Ashok, 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.

3. Basu, D. 1971. An essay on the Logical Foundations of Survey Sampling, Part 1, *Foundations of Statistical Inference*, 203-242. Toronto: Holt, Rinehart and Winston.

4. Brewer, K.R.W. 1963. A Model of Systematic Sampling with Unequal Probabilities. *Australian Journal of Statistics*, 5, 5-12.

5. Cassel, C.M., Särndal, C.E. and Wretman, J.H. 1977. *Foundations of Inference in Survey Sampling*. New York: Wiley.

6. Chaudhuri, A. 1988. Optimality of Sampling Strategies. *Handbook of Statistics Volume 6*, 47-96. Amsterdam: North Holland

7. Chaudhuri, A. and Arnab, R. 1979. On the Relative Efficiencies of Sampling under a Super Population Model. *Sankhyā*, C, 41, 40-53.

8. Chaudhuri, A. and Stenger, H. 1992. Survey *Sampling: Theory and Methods*. New York: Marcel Dekker.

9. Cochran, W.G. 1977. *Sampling Techniques*, 3rd ed. New York: Wiley.

10. Durbin, J. 1967. Design of Multi-Stage Surveys for the Estimation of Sampling Errors. *Applied Statistics*, 16, 2, 152-164.

11. Godambe, V.P. 1955. A unified Theory of Sampling from Finite Populations. *Journal of the Royal Statistical Society*, B, 17, 269-278.

12. Godambe, V.P. 1960. An optimum property of Regular Maximum Likelihood Estimation. *The Annals of Mathematical Statistics*, 1208-1211.

13. Godambe, V.P. and Joshi, V.M. 1965. Admissibility and Bayes Estimation in Sampling Finite Populations I. *The Annals of Mathematical Statistics*, 1707-1722.

14. Goodman, R. and Kish, l. 1950. Controlled Selection-A Technique in Probability Sampling. *Journal of the American Statistical Association*, 45, 350-372.

15. Hansen, M.H., Dalenius, T. and Tepping, B.J. 1985. The development of Sample Surveys of Finite Populations. *A celebration of Statistics: The ISI Centenary Volume*. New York: Springer.

16. Hansen, M.H. and Hurwitz, W.H. 1943. On the Theory of Sampling from Finite Populations. *The Annals of Mathematical Statistics*, 333-362.

17. Hansen, M.H., Madow, W.G. and Tepping, B.J. 1983. An evaluation of Model-Dependent and Probability-Sampling Inferences in sampling surveys. *Journal of the American Statistical Association*, 78, 776-793.

18. Hanurav, T.V. 1965. Optimal Sampling Strategies. *PhD thesis submitted to the Indian Statistical Institute.*

19. Hanurav, T.V. 1966. Some aspects of Unified Sampling Theory. *Sankhyā*, A, 28, 175-204.

20. Hanurav, T.V. 1967. Optimum Utilization of Auxiliary Information: π ps Sampling of Two units from a Stratum. *Journal of the Royal Statistical Society*, B, 29, 374-391.

21. Hartley, H.O. and Rao, J.N.K 1978. Estimation of Non-Sampling Variance Components in Sample Surveys. *Survey Sampling and Measuremen*t, 35-43.

22. Horvitz, D.G. and Thompson, D.G. 1952. A Generalization of Sampling Without Replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-686.

23. Johnson, N.L. and Smith, H. Jnr. 1969. *New Developments in Survey Sampling*. New York: Wiley Inter-Science.

24. Krishnaiah, P.R. and Rao, C.R. 1988. *Handbook of Statistics, Volume 6.* Amsterdam: North-Holland.

25. Lohr, S.L. 1999. *Sampling: Design and Analysis*. USA: Brooks/Cole Publishing Co.

26. Midzuno, H. 1952. On the Sampling System with Probabilities Proportional to sum of sizes. *Annals of the Institute of Statistical Mathematics*, 3, 99-107.

27. Murthy, M.N. 1957. Ordered and Unordered Estimators in Sampling without Replacement. Sankhyā, A, 18, 379-390.

28. Raj, D. 1956. A note on the Determination of Optimum Probabilities in Sampling Without Replacement. *Sankhyā*, 17, 197-200.

29. Rao, J.N.K. 1994. Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. *Journal of Official Statistics*, 10, 2, 153-165.

30. .Rao, J.N.K., Hartley, H.O. and Cochran, W.G. 1962. On a simple procedure of Unequal Probability Sampling without Replacement. *Journal of the Royal Statistical Society*, B, 24, 482-491.

31. Rao, T.J. 1967. On the Choice of a Strategy for the Ratio Method of Estimation. *Journal of the Royal Statistical Society*, B, 29, 392-397.

32. Royall, R.M. 1970. On Finite Population theory under certain Linear Regression Models. *Biometrika*, 57, 377-387.

33. Sen, A.R. 1953. On the Estimator of the Variance in Sampling with Varying Probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5:2, 119-127.

34. Valliant, R., Dorfman, A.H. and Royall, R.M. 2000. *Finite Population Sampling and Inference. A Prediction Approach*. New York: John Wiley.

35. Yates, F. and Grundy, P.M. 1953. Selections Without Replacement from within strata with Probability Proportional to Size. *Journal of the Royal Statistical Society*, B, 15, 253-261.