Fundamentals of Survey Sampling

Introduction
The purpose of sampling theory is to develop methods of sample selection from a finite population and of estimation that provide estimates of the unknown population parameters, generally population total or population mean, which are precise enough for our purpose. Survey samples can broadly be categorized into two types: probability samples and non-probability samples. Surveys based on probability samples are capable of providing mathematically sound statistical inferences about a larger target population. Inferences from probability-based surveys may, however, suffer from many types of bias.

There is no way of measuring the bias or sampling errors of estimators in surveys which are not based on probability sampling. Surveys based on non-probability samples are not externally valid. They can only be said to be representative of sampling units that have actually been observed. Henceforth, a sample survey would always mean one wherein sampling units have been selected by probability sampling mechanism, unless otherwise stated.

Sample survey methods, based on probability sampling, have more or less replaced complete survey (or census) methods on account of several well known advantages. It is well recognized that the introduction of probability sampling approach has played an important role in the development of survey techniques. The concept of representativeness through probability sampling techniques introduced by Neyman (1934) provided a sound base to the survey approach of data collection. One of the salient features of probability sampling is that besides providing an estimate of the population parameter, it also provides an idea about the precision of the estimate (sampling error). Throughout this lecture, the attention would be restricted to sample surveys and not the complete survey. For a detailed exposition of the concepts of sample survey, reference may be made to the texts of Cochran (1977), Desraj and Chandok (1998), Murthy (1977), Sukhatme et al. (1984), Mukhopadhyay (1998).

Population, sample, estimator
A finite population is a collection of known number $N$ of distinct and identifiable sampling units. If $U$’s denote the sampling units, the population of size $N$ may be represented by the set $U = \{U_1, U_2, \ldots, U_i, \ldots, U_N \}$. The study variable is denoted by $y$ having value $Y_i$ on unit
\(i; i = 1, 2, \ldots, N\). We may represent by \(Y = (Y_1, Y_2, \ldots, Y_i, \ldots, Y_N)\) an \(N\)-component vector of the values of the study variable \(Y\) for the \(N\) population units. The vector \(Y\) is assumed fixed though unknown. Sometimes auxiliary information is also available on some other characteristic \(x\) related with the study variable \(Y\). The auxiliary information is generally available for all the population units. We may represent by \(X = (X_1, X_2, \ldots, X_i, \ldots, X_N)\) an \(N\)-component vector of the values of the auxiliary variable \(X\) for the \(N\) population units. The total \(X = X_1 + X_2 + \cdots + X_i + \cdots + X_N\) is generally known.

A list of all the sampling units in the population along with their identity is known as sampling frame. The sampling frame is a basic requirement for sampling from finite populations. It is assumed that the sampling frame is available and it is perfect in the sense that it is free from under or over coverage and duplication.

The probability selection procedure selects the units from \(U\) with probability \(P_i, i \in U\). We shall denote by \(P = (P_1, P_2, \ldots, P_i, \ldots, P_N)\) an \(N\)-component vector of the initial selection probabilities of the units such that \(P^\prime 1 = 1\). Generally \(P = g(n, N, X)\); e.g., \(P_i = 1/N\ \forall \ i \in U\); or \(P_i = n/N\ \forall \ i = 1, 2, \ldots, k, k = N/n\); or \(P_i = X_i/(X_1 + X_2 + \ldots + X_N), \forall \ i \in U\).

A nonempty set \(\{s : s \subseteq U\}\), obtained by using probability selection procedure \(P\), is called an unordered sample. The cardinality of \(s\) is \(n\), which is also known as the (fixed) sample size. However, there shall be occasions wherein we shall discuss the equal probability with replacement and unequal probability with replacement sampling schemes. In such cases the sample size is not fixed. So barring these exceptions, we shall generally assume a fixed sample size \(n\) throughout. A set of all possible samples is called sample space \(S\). While using probability selection procedures, the sample \(s\) may be drawn either with or without replacement of units. In case of with replacement sampling, the cardinality of \(S\) is \(\eta = N^n\), and the probability is very strong that the sample selected may contain a unit more than once. In without replacement sampling the cardinality of \(S\) is \(\nu = \binom{N}{n}\), and the probability that the sample selected may contain a unit more than once is zero. Throughout, it is assumed that the probability sampling is without replacement of units unless specified otherwise.
Given a probability selection procedure $P$ which describes the probability of selection of units one by one, we define the probability of selection of a sample $s$ as $p(s) = g(P_i) : i \in s$, $s \in S$. We also denote by $p = \{p(1), p(2), \ldots, p(s), \ldots, p(n)\}$ a $v$-component vector of selection probabilities of the samples. Obviously, $p(s) \geq 0$ and $p'1 = 1$. It is well known that given a unit by unit selection procedure, there exist a unique mass selection procedure; the converse is also true.

After the sample is selected, data are collected from the sampled units. Let $y_i$ be the value of study variable on the $i^{th}$ unit selected in the sample $s$, $i \in s$ and $s \in S$. We shall denote by $y = (y_1, y_2, \ldots, y_t, \ldots, y_n)'$ an $n$-component vector of the sampled observations. It is assumed here that the observation vector $y$ is measured without error and its elements are the true values of the sampled units.

The problem in sample surveys is to estimate some unknown population parameter $\theta = f(Y)$ or $\theta = f_1(Y, X)$. We shall focus on the estimation of population total, $\theta = Y'1 = \sum_{i \in U} Y_i$, or population mean $\theta = \bar{Y} = N^{-1} Y'1 = N^{-1} \sum_{i \in U} Y_i$. An estimator $e$ for a given sample $s$ is a function such that its value depends on $y_i$, $i \in s$. In general $e = h(y, X)$ and the functional form $h(\cdot, \cdot)$ would also depend upon the functional form of $\theta$, besides being a function of the sampling design. We can also write $e_s = h\{y, p(s)\}$.

A sampling design is defined as

$$d = \{[s, p(s)] : s \in S\}. \quad (1)$$

Further $\sum_{s \in S} p(s) = 1. \quad (2)$

We shall denote by $D = \{d\}$ a class of sampling designs.

$S$ is also called the support of the sampling plan and $v = N C_n$ is called the support size. A sampling plan is said to be a fixed-size sampling plan if whenever $p(s) > 0$, the
corresponding subsets of units are composed of the same number of units. We shall restrict
the discussion to fixed size samples only and sample size would always mean fixed size
sample.

The triplet \((S, p, e_s)\) is called the \textit{sampling strategy}.

A familiarity with the expectation and variance operators is assumed in the sequel. An
estimator \(e_s\) is said to be \textit{unbiased} for estimation of population parameter \(\theta\) if
\(E_d(e_s) = \theta\) with respect to a sampling design \(d\), where \(E_d\) denotes the expectation operator. The \textit{bias} of
an estimator \(e_s\) for estimating \(\theta\), with respect to a sampling design \(d\), is \(B_d(e_s) = E_d(e_s) - \theta\).
\textit{Variance} of an unbiased estimator \(e_s\) for \(\theta\), with respect to sampling design \(d\), is
\(V_d(e_s) = E_d[(e_s - E_d(e_s))^2] = E_d(e_s^2) - \theta^2\). The \textit{mean square error} of a biased
estimator \(e_s\) for \(\theta\) is given by \(MSE_d(e_s) = E_d(e_s - \theta)^2 = E_d(e_s^2 - E_d(e_s) + E_d(e_s) - \theta)^2 = V_d(e) + B_d(e)^2\).

\textbf{Implementation of sampling plans}

Consider again a finite population of \(N\) distinct and identifiable units. The problem is to
estimate some population parameter \(\theta\) using a sample of size \(n\) drawn without replacement of
units using some pre-defined sampling plan \(p = \{p(s); s \in S\}\). There are \(v\) samples in the
sample space \(S\). The sampling distribution is given in the Table below:

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Sample (s)</th>
<th>Probability of selection (p(s))</th>
<th>Cumulative Sum of (p(s))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(s_1)</td>
<td>(p(s_1))</td>
<td>(p(s_1))</td>
</tr>
<tr>
<td>2</td>
<td>(s_2)</td>
<td>(p(s_2))</td>
<td>(p(s_1) + p(s_2))</td>
</tr>
<tr>
<td>3</td>
<td>(s_3)</td>
<td>(p(s_3))</td>
<td>(p(s_1) + p(s_2) + p(s_3))</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>(k - 1)</td>
<td>(s_{k-1})</td>
<td>(p(s_{k-1}))</td>
<td>(p(s_1) + p(s_2) + p(s_3) + \ldots + p(s_{k-1}))</td>
</tr>
<tr>
<td>(k)</td>
<td>(s_k)</td>
<td>(p(s_k))</td>
<td>(p(s_1) + p(s_2) + p(s_3) + \ldots + p(s_k))</td>
</tr>
<tr>
<td>(k + 1)</td>
<td>(s_{k+1})</td>
<td>(p(s_{k+1}))</td>
<td>(p(s_1) + p(s_2) + p(s_3) + \ldots + p(s_{k+1}))</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>(v)</td>
<td>(s_v)</td>
<td>(p(s_v))</td>
<td>(p(s_1) + p(s_2) + p(s_3) + \ldots + p(s_v) = 1)</td>
</tr>
</tbody>
</table>
Now, draw a random number between 0 and 1 (i.e. a uniform variate). Let the random number drawn be \( R \).

Then the \( k^{\text{th}} \) sample would be selected if
\[
p(s_1) + p(s_2) + p(s_3) + \ldots + p(s_{k-1}) < R \leq p(s_1) + p(s_2) + p(s_3) + \ldots + p(s_k).
\]

**Selection probability and inclusion probability**

Suppose a sample of fixed size \( n \) is selected without replacement of units from a population containing \( N \) distinct and identifiable units by drawing units one by one in a total of \( n \) draws.

Let \( p_i^{(r)} \) denote the probability of selecting the \( i^{\text{th}} \) unit at the \( r^{\text{th}} \) draw, \( i = 1, 2, \ldots, N; r = 1, 2, \ldots, n \). \( p_i^{(r)} \) satisfies the following conditions:

(i) \( \sum_{i=1}^{N} p_i^{(r)} = 1 \) and (ii) \( 0 < p_i^{(r)} < 1 \) for any \( 1 \leq r \leq n \).

The **first-order inclusion probability** for unit \( i \) is the probability that unit \( i \) is included in a sample of size \( n \) and is given by
\[
\pi_i = \sum_{s \ni i} p(s). \tag{3}
\]

The **second-order inclusion probability** for units \( i \) and \( j \) is defined as the probability that the two units \( i \) and \( j \) are included in a sample of size \( n \) and is given by
\[
\pi_{ij} = \sum_{s \ni i, j} p(s). \tag{4}
\]

For any sampling design \( d \), the inclusion probabilities satisfy following conditions.

i) \( \sum_{i=1}^{N} \pi_i = n \); ii) \( \sum_{j=1}^{N} \pi_{ij} = (n-1)\pi_i, \forall i = 1, 2, \ldots, N \); iii) \( \sum_{i=1}^{N} \sum_{j=1}^{N} \pi_{ij} = n(n-1) \).

**Horvitz-Thompson estimator**

For a given population of \( N \) units, define

Population Total, \( Y = \sum_{i=1}^{N} Y_i \), Population Variance, \( \sigma^2 = N^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \),
and Population Mean, $\bar{Y} = N^{-1} \sum_{i=1}^{N} Y_i$. \hfill (5)

For an observed sample, $s$, obtained using a sampling design $\{s, p(s); S \in S\}$, the Horvitz-Thompson estimator for the population total is defined as

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}. \hfill (6)$$

Provided that all first-order inclusion probabilities of units in the population are different from zero, (6) is an unbiased estimator of the population total.

The Sen-Yates-Grundy (1953) form of variance of the Horvitz-Thompson estimator of the population total is

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \sum_{i < j=1}^{N} (\pi_i \pi_j - \pi_{ij}) \left( \frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2. \hfill (7)$$

Provided that the second-order inclusion probabilities for all pairs of units in the population are non-zero, an unbiased estimator of variance is given by

$$V(\hat{Y}_{HT}) = \sum_{i \in s} \sum_{i < j \in s} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \hfill (8)$$

Clearly, a sufficient condition for non-negativity of variance estimator (8) is that $\pi_i \pi_j \geq \pi_{ij}$, $\forall i \neq j$.

**Equal and unequal probability sampling**

Probability sampling procedures can broadly be classified into three broad categories, viz., (i) simple random sampling, (ii) systematic sampling and (iii) unequal probability sampling.
**Simple random sampling**

Simple random sampling (SRS) is the simplest method of selecting a sample of \( n \) units out of \( N \) units by drawing units one by one with or without replacement.

**Simple random sampling with replacement**

Simple random sampling with replacement (SRSWR) is a method of drawing \( n \) units from \( N \) units in the population such that at every draw, each unit in the population has the same chance of being selected and draws are made with replacement of selected units. For SRSWR scheme, \( P_i = \frac{1}{N} \quad \forall \, i = 1, 2, \ldots, N \). Further, \( P_i^{(r)} = P_i^{(r)} = \frac{1}{N} \quad \forall \, i \in U; \, r = 1, 2, \ldots, n; \, s \in S \).

The sample space \( S \) under SRSWR consists of \( N^n \) possible samples. Probability of selecting a sample of size \( n \) under SRSWR is

\[
p(s) = \frac{1}{N} \cdot \frac{1}{N} \cdots \frac{1}{N} = \frac{1}{N^n}.
\]

The first and second order inclusion probabilities under SRSWR are given by,

\[
\pi_i = 1 - \left(1 - \frac{1}{N}\right)^n, \quad (i = 1, 2, \ldots, N) \quad \text{and} \quad \pi_{ij} = 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n, \quad (i \neq j = 1, 2, \ldots, N).
\]

An unbiased estimator of population total \( Y \) under SRSWR is given by

\[
\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i = N\bar{y},
\]

where \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \) is sample mean of \( y \).

The variance of the estimator (9) under SRSWR is

\[
V(\hat{Y}) = N^2 \frac{\sigma^2}{n},
\]

where \( \sigma^2 \) is as defined in (5).

An unbiased estimator of (10) under SRSWR is given by

\[
\hat{V}(\hat{Y}) = N^2 \frac{s^2}{n},
\]

where \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \).
Simple random sampling without replacement

SRSWR has the drawback that one or more sampling unit may occur more than once in a sample. Repeated units provide no extra information for estimation of parameters. Simple random sampling without replacement (SRSWOR) is free from such loss of information and is thus preferred over SRSWR. The SRSWOR is a method of drawing \( n \) units from \( N \) units in the population such that at every draw, each unit in the population has the same chance of being selected and draws are made without replacement of selected units. The sample space \( S \) under SRSWOR consists of \( \binom{N}{n} \) possible samples.

The selection probability of a unit at any given draw in case of SRSWOR is \( \frac{1}{N} \), i.e.,

\[
P_i^{(r)} = P_i^{(s)} = \frac{1}{N} \quad \forall \; i \in U; \; r = 1, 2, \cdots, n; \; s \in S.
\]

Probability of selecting a sample of size \( n \) under SRSWOR is \( p(s) = \frac{1}{\binom{N}{n}} \).

Under SRSWOR, \( \pi_i = \frac{n}{N} \), \((i = 1, 2, \ldots, N)\) and \( \pi_{ij} = \frac{n(n-1)}{N(N-1)} \), \((i \neq j = 1, 2, \ldots, N)\).

An unbiased estimator of population total \( Y \) under SRSWOR is

\[
\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i = N\bar{y}.
\]

The variance of the estimator (12) under SRSWOR is given by

\[
V(\hat{Y}) = \frac{N^2(n-n)\sigma^2}{(N-1)n} = \frac{N^2(n-n)S^2}{Nn},
\]

where \( \sigma^2 \) is as defined in (5) and \( N\sigma^2 = (N-1)S^2 \).

An unbiased estimator of (13) under SRSWOR is given by

\[
\hat{V}(\hat{Y}) = \frac{N(N-n)s^2}{n},
\]

where \( (n-1)s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \).
It may be seen that Horvitz-Thompson estimator of population total $Y$ under SRSWOR is identical with the estimator in (12).

The variance of Horvitz-Thompson estimator of population total $Y$ under SRSWOR is
\[ V(\hat{Y}_{HT}) = \frac{(N-n)N^2}{(N-1)n} \sigma^2, \quad (15) \]
which is same as (13).

An unbiased estimator of Sen-Yates-Grundy form of variance of Horvitz-Thompson estimator of population total $Y$ under SRSWOR is
\[ \hat{V}(\hat{Y}_{HT}) = \sum_{i<s} \sum_{i<j<s} \frac{\pi_i\pi_j - \pi_0}{\pi_j} \left( \frac{y_i - y_j}{\pi_i - \pi_j} \right)^2 = \frac{N(N-n)}{n} s^2, \quad (16) \]
which is same as (14).

**Systematic sampling**

Systematic sampling is an operationally very convenient sampling technique in which only the first unit is selected randomly and the rest of the units get selected automatically according to some pre-specified pattern. If $N$ units of the population are labeled as 1, 2, ..., $N$ in some order. Let $N = nt$, where $n$ is sample size and $t$ is a positive integer. Then under systematic sampling, a random number $u$ less than or equal to $t$ is selected. The sample then consists of the unit $u$ and every $t^{th}$ unit thereafter and thus the sample is \{u, u + t, ..., u + (n - 1)t\}. It may be seen there are $t$ systematic samples and each sample has the probability of selection $1/t$. Also, $\pi_i = \frac{1}{t} = \frac{n}{N} \forall i$ and $\pi_j = \frac{1}{t} = \frac{n}{N} \forall j = i + t, i + 2t, ..., i + (n - 1)t$ and $= 0$ for other $j$’s.

For $N \neq nt$, Lahiri (1954) suggested circular systematic sampling in which $t$ is taken as the integer nearest to $\frac{N}{n}$. A random number $u$ is chosen from 1 to $N$ and the $u$th unit is selected and thereafter every $t^{th}$ unit is chosen in a circular manner till a sample of $n$ units is selected.

When $N = nt$, an unbiased estimator of population total is given by
\[
\hat{Y}_{\text{sys}} = \frac{N}{n} \sum_{j=1}^{n} y_{ij}
\]  
(17)

where \( y_{ij} \) denotes the observation on the \( j \)-th \( (j = 1,2,...,n) \) unit in the \( i \)-th \( (i = 1,2,...,t) \) systematic sample.

However, the estimator (17) is not unbiased when \( N \neq nt \). In general, an unbiased estimator of population total \( Y \) is

\[
\hat{Y}_{\text{sys}}^* = \frac{N}{n} \sum_{j=1}^{n'} y_{ij}
\]  
(18)

where \( n' \) is the size of the selected sample.

The variance of the estimator (17) is given by

\[
V(\hat{Y}_{\text{sys}}) = \frac{1}{t} \sum_{t=1}^{t} \left( \hat{Y}_{\text{sys}} - Y \right)^2.
\]  
(19)

Unbiased estimation of variance at (19) is not possible. To see this, note

\[
V(\hat{Y}_{\text{sys}}) = E(\hat{Y}_{\text{sys}}^2) - Y^2 = E(\hat{Y}_{\text{sys}}^2) - \sum_{u=1}^{N} Y_u^2 - \sum_{u \neq u'} Y_u Y_{u'}.
\]

Since all possible pairs \( (u,u') \) do not appear in the sample space, third term cannot be estimated unbiasedly. However, an approximate variance estimator due to Cochran (1946) is

\[
V(\hat{Y}_{\text{sys}}) = \frac{N(N-n)}{n(n-1)} \left( \sum_{j=1}^{n} y_{ij}^2 - \frac{n \hat{Y}_{\text{sys}}^2}{N^2} \right).
\]  
(20)

Another approximate estimate of variance due to Yates (1948) is given by

\[
V(\hat{Y}_{\text{sys}}) = \frac{N(N-n)}{2n(n-1)} \sum_{j=1}^{n} (y_{ij} - y_{ij+1})^2.
\]  
(21)

Both the estimators (20) and (21) are biased and hence should be used with caution.

**Unequal probability sampling**

In sampling from finite population, often the value of some auxiliary variable, closely related to the main characteristic of interest, is available for all the units of the population. This normalized value of the auxiliary variable may be taken as a measure of the size of a unit. For example, in agricultural surveys, area under crop may be taken as a size measure of farms for
estimating the yield of crops. In such situations sampling the units with probability proportional to size measure with replacement or without replacement may be used in place of SRSWWR or SRSWOR. Since the units with larger sizes are expected to have bigger total of \( Y \), it is expected that probability proportional to size measure sampling procedures will be more efficient than SRSWOR or SRSWR. Note that sampling units have unequal probability of selection. In the sequel we discuss some common unequal probability sampling procedures.

**Probability proportional to size with replacement**

In probability proportional to size with replacement (PPSWR) sampling, the units are selected with probabilities proportional to some measure of their size and with replacement of units. Let \( P_i = \frac{X_i}{\sum_{i=1}^{N} X_i} \) denote the normalized size measure of the \( i^{th} \) \((i = 1,2,...,N)\) unit, where \( X_i \) denotes the size measure of the \( i^{th} \) unit. In PPSWR, the \( i^{th} \) unit is selected with probability \( P_i \).

Note that \( \sum_{i=1}^{N} P_i = 1 \). The first and second order inclusion probabilities are given by \( \pi_i = 1 - (1 - P_i)^n; i = 1,2,...,N \) and \( \pi_{ij} = 1 - (1 - P_i)^n - (1 - P_j)^n + (1 - P_i - P_j)^n; i \neq j = 1,2,...,N \).

A PPSWR sample may be selected in two ways namely cumulative total method and Lahiri’s method. In cumulative total method, cumulative total of sizes of the units is made. Let \( T_i = \sum_{j=1}^{i} X_j \). A random number between 1 and \( T_N = \sum_{j=1}^{N} X_j \) is drawn. Let the number drawn be \( r \). If \( T_{i-1} < r \leq T_i \), then the unit \( i \) is selected. This method is repeated \( n \) times for selecting a sample of size \( n \). It may be seen that the method is cumbersome when population size is large. Lahiri’s method is used when number of units is very large. The method involves drawing two random numbers \( i \) and \( r \); first random number \( i \) between 1 and \( N \) and second random number \( r \) between 1 and \( M \), where \( M = \max(X_i); i \in U \). Select the unit \( i \) if \( r \leq X_i \). Otherwise reject the pair of random numbers \( i \) and \( r \) and draw a fresh pair of random numbers. Repeat the trial till a unit is selected. This whole process is repeated \( n \) times to select a sample of size \( n \).
An unbiased estimator of population total $Y$ under PPSWR sampling is

$$
\hat{Y} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{P_i}.
$$

Under PPSWR sampling, variance of the estimator (22) of population total $Y$ is given by

$$
V(\hat{Y}) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i}{P_i} - Y \right)^2.
$$

An unbiased estimator of (23) is

$$
\hat{V}(\hat{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left( \frac{y_i}{P_i} - \hat{Y} \right)^2.
$$

**Unequal probability sampling without replacement**

Most of the unequal probability sampling without replacement designs/schemes available in the literature are based on Horvitz-Thompson estimator. Some desirable properties of a sampling design based on Horvitz-Thompson estimator, listed by Hanurav (1967), are

i) $\pi_i = nP_i \forall i \in U$

ii) $\pi_i > 0 \forall i \in U; \pi_{ij} > 0, \forall i \neq j \in U$

iii) $\pi_{ij} \leq \pi_i, \pi_j, \forall i \neq j = 1,2,...,N$

iv) $\phi_{ij} = \frac{\pi_{ij}}{\pi_i\pi_j} > \varphi, \varphi$ is not too close to zero.

Sampling plans satisfying property (i) are called Inclusion Probability Proportional to Size (IPPS) sampling plans. If the choice of the auxiliary variable is such that the study variable is proportional to the auxiliary variable, then the variance of the Horvitz-Thompson estimator of the population total will be zero. However, a near proportionality will ensure that the variance of the Horvitz-Thompson estimator of the population total is very small. In that sense the IPPS schemes (or $\pi PS$) sampling schemes have been studied extensively in the literature. Property (ii) is required for existence of an unbiased estimator of the population total and an unbiased variance estimator of the population total. Property (iii) ensures non-negativity of the variance estimator while property (iv) guarantees stability of the variance estimator.
In the sequel we give a description of some of the IPPS sampling schemes.

**Inclusion probability proportional to size sampling designs**

Inclusion probability proportional to size (IPPS) sampling is a sampling procedure in which units are selected without replacement and for which \( \pi_i \), the probability of including the \( i^{th} \) unit in a sample of size \( n \) is \( nP_i \), where \( P_i \) is the initial probability of selecting \( i^{th} \) unit in the population. An estimator which is commonly used to estimate the population total with IPPS sampling procedures is the well known Horvitz-Thompson (1952) estimator. Narain (1951) showed that a necessary condition for Horvitz-Thompson estimator of population total under IPPS to be better than estimator (22) under PPSWR is that

\[
\frac{\pi_{ij}}{\pi_i \pi_j} \leq \frac{2(n-1)}{n} \forall i \neq j \in U.
\]

A sufficient condition for Horvitz-Thompson estimator of population total under IPPS to be better than estimator (22) under PPSWR is that

\[
\frac{\pi_{ij}}{\pi_i \pi_j} > \frac{n-1}{n} \forall i \neq j \in U.
\]

There are a number of IPPS sampling designs available in literature. Initially the focus was on obtaining IPPS sampling schemes for sample size \( n = 2 \). For a general sample size, \( n > 2 \) not many IPPS schemes are available in the literature. Some important IPPS schemes for \( n = 2 \) are: modified Midzuno-Sen (1952) strategy, Yates-Grundy (1953) sampling design, Brewer’s (1963) sampling design, Durbin’s (1967) ungrouped procedure and Rao’s (1965) rejective procedure. IPPS sampling plans for \( n \geq 2 \) are Fellegi’s (1963) scheme, Sampford’s (1967) scheme, Tille’s method, Midzuno-Sen (1952), Rao-Hartley-Cochran (1962), and Gupta et al. (1982, 1984).

We now describe briefly some important sampling plans for sample size \( n \geq 2 \).

**Sampford’s (1967) IPPS scheme**

The most important sampling scheme available in the literature for obtaining an IPPS sample of size \( n \geq 2 \) is due to Sampford (1967). This sampling scheme provides a non-negative variance estimator, is better than PPSWR scheme and provides stable variance estimator. We describe the scheme below:

Let \( \gamma_i = \frac{P_i}{1 - P_i} \), \( S(m) = \{i_1, i_2, ..., i_m : \text{i}_j \text{'s are all distinct, } j = 1, 2, ..., m\} \).
\[ L_m = \sum_{s(m)} \gamma_i \gamma_j \cdots \gamma_k; \quad (1 \leq m \leq N) \text{ and } L_0 = 1. \]

A sample \( s(i_1, i_2, \ldots, i_n) \) is selected with probability

\[ p(s) = K_n \sum_{u=1}^{n} P_{u} \sum_{v(u) \neq 1}^{n} \gamma_{i_v} = nK_n \gamma_{i_1} \gamma_{i_2} \cdots \gamma_{i_n} \left( 1 - \sum_{u=1}^{n} P_{u} \right), \quad (25) \]

where

\[ K_n = \left( \sum_{i=1}^{n} tL_{n-i} \right)^{-1}. \]

The above probability of selection can be realized by drawing units one by one.

The above sampling plan ensures \( \pi_i = nP_i \) for \( i = 1, 2, \ldots, N \). For proofs, a reference may be made to Sampford (1967).

Sampford’s IPPS sampling plan ensures \( \pi_{ij} > 0 \quad \forall \ i \neq j = 1, 2, \ldots, N \) and \( \pi_{ij} < \pi_i \pi_j \quad \forall \ i \neq j = 1, 2, \ldots, N \) (sufficient condition for non-negativity of variance estimator).

**Midzuno-Sen (1952) sampling strategy**

Midzuno-Sen (1952) sampling strategy is one of the simplest sampling designs. Under this sampling design, the first unit is chosen with probability \( P_i \) and the remaining \( (n-1) \) units are selected from the remaining \( (N-1) \) population units by SRSWOR.

Under Midzuno-Sen (1952) sampling,

\[ \pi_i = \frac{N-n}{N-1} P_i + \frac{n-1}{N-1} \quad \forall \ i \in U \]

\[ \pi_{ij} = \frac{(n-1)(N-n)}{(N-1)(N-2)} \left( P_i + P_j \right) + \frac{(n-1)(n-2)}{(N-1)(N-2)}, \quad \forall \ i \neq j \in U. \quad (26) \]

It is easy to see that if the first unit is selected with revised probability of selection

\[ P_i^* = \frac{N-1}{N-n} nP_i - \frac{n-1}{N-n} \]

and the remaining \( (n-1) \) units are selected by SRSWOR from the remaining \( (N-1) \) population units then the plan reduces to an IPPS sampling plan. Since,
probability of selection of every population unit should be different from zero, to achieve an IPPS sampling plan with such revised probabilities requires the condition \( \frac{n-1}{n(N-1)} \leq P_t \leq \frac{1}{n} \) to be satisfied.

Now, we describe another very important strategy due to Rao-Hartley-Cochran (1962).

**Rao-Hartley-Cochran (RHC) sampling strategy**

Rao, Hartley and Cochran (1962) proposed a very simple sampling strategy. First the population is randomly divided into \( n \) groups \( G_1, G_2, \ldots, G_n \), of sizes \( N_1, N_2, \ldots, N_n \) units such that \( N = N_1 + N_2 + \ldots + N_n \). Let \( P_{tg} \) denote the initial selection probability of unit \( t \) from the \( g \)th group, \( t = 1,2,\ldots,N_g \) and \( g = 1,2,\ldots,n \). Further, \( \sum_{t=1}^{N_g} P_{tg} = \Pi_g \) and \( \sum_{g=1}^{n} \Pi_g = 1 \). From the \( g \)th group, \( t \)th unit is selected with probability \( \frac{P_{tg}}{\Pi_g} \), \( g = 1,2,\ldots,n \).

An unbiased estimator of population total \( Y \) under RHC sampling is

\[
\hat{Y}_{RHC} = \sum_{g=1}^{n} \frac{Y_{tg}}{P_{tg}} \Pi_g .
\]  

(27)

The variance of the estimator (27) of population total \( Y \) under RHC sampling is

\[
V(\hat{Y}_{RHC}) = \frac{n}{N(N-1)} \left( \sum_{g=1}^{n} \frac{N_g^2 - N}{n} - \frac{Y^2}{n} \right).
\]  

(28)

In case of equal group sizes i.e., \( N_1 = N_2 = \ldots = N_n = \frac{N}{n} \),

\[
V(\hat{Y}_{RHC}) = \left( 1 - \frac{n-1}{N-1} \right) \left( \sum_{i=1}^{N} \frac{Y_i^2}{nP_i} - \frac{Y^2}{n} \right).
\]

An unbiased estimator of (28) is given by

\[
\hat{V}(\hat{Y}_{RHC}) = \frac{\sum_{g=1}^{n} \frac{N_g^2 - N}{n}}{N^2 - \sum_{g=1}^{n} N_g^2} \sum_{g=1}^{n} \Pi_g \left( \frac{Y_{tg}}{P_{tg}} - \frac{\hat{Y}_{RHC}}{g} \right)^2 .
\]
**Classes of linear estimators**

Horvitz-Thompson (1952) described various sub-classes of estimators. Some sub-classes of estimators, for example, are the following:

(a) Considering the order of appearance of the elements, the first class of estimators is of the form 
\[ e_s = \sum_{r=1}^{n} \alpha_r y_r \]
where the weight \( \alpha_r (r = 1, 2, \ldots, n) \) is a function of the order of the drawing and is the weight to be attached to the element selected at the \( r \)th draw, \( \alpha_r (r = 1, 2, \ldots, n) \) being defined in advance. Here the subscripts identifying the elements have been dropped from the \( y \)'s, meaning thereby that \( y_r \) is the value of the characteristic (or the study variable) corresponding to the element selected at the \( r \)th draw. This estimator is called as the \( T_1 \) class of estimators.

(b) Considering the presence or absence of an element in the sample, the second class of estimators is of the form 
\[ e_s = \sum_{i=1}^{n} \beta_i y_i \]
where \( \beta_i \) is the weight to be attached to the \( i \)th element whenever it appears in the sample and is defined in advance for all \( i = 1, 2, \ldots, n \). This estimator is called as the \( T_2 \) class of estimators.

(c) Considering the sample obtained as one of the set of all possible distinct samples, the third class of estimators is of the form 
\[ e_s = \gamma_s \sum_{r=1}^{n} y_r \]
where \( \gamma_s \) is the weight to be attached to the \( s \)th sample whenever it is selected, and is defined for all \( s \in S \). This estimator is called as the \( T_3 \) class of estimators.

It is noteworthy that this classification of the estimators is not exhaustive. Koop (1957) described seven classes of estimators, which included the three estimators of Horvitz and Thompson (1952). Godambe (1955) described the most general class of estimators and referred as to the generalized linear estimators.

Suppose that the parameter of interest is the population total, i.e., \( \theta = \sum_{i \in U} y_i \). As mentioned above, the population size is \( N \) and the sample size is \( n \).

In case of SRSWOR, the estimator of \( \theta \) is 
\[ \frac{N}{n} \sum_{r \in s} y_r \] . This estimator belongs to \( T_2 \) class of estimators. Similarly, in case of SRSWR, the usual estimator of \( \theta \) is 
\[ \frac{N}{n} \sum_{i=1}^{n} y_i \] . This estimator belongs to \( T_1 \) class of estimators. On the other hand, for SRSWR, an unbiased
estimator of $\theta$ is $e_s = \sum_{r \in s} y_r \left[ 1 - \left(1 - \frac{1}{N} \right)^n \right]$. This estimator belongs to $T_2$ class of estimators. Similarly, in case of stratified sampling with SRSWOR in each stratum, the usual unbiased estimator of $\theta$ is $e_s = \frac{\sum_{h=1}^{L} N_h}{n} \sum_{r \in s_h} y_r$. This estimator also belongs to $T_2$ class of estimators. Here $L$ denotes the number of strata in which the population is divided, $N_h$ is the size of the $h^{th}$ stratum, $\sum_{h=1}^{L} N_h = N$ and $n_h$ is the size of the sample $s_h$ drawn from the $h^{th}$ stratum, $\sum_{h=1}^{L} n_h = n$. Further, let $x$ be an auxiliary variable, correlated with the study variable $y$. Then the usual ratio estimator of $\theta$ is $e_s = \frac{\sum_{r \in s} y_r}{\sum_{r \in s} x_r} \sum_{i \in s} x_i$. This estimator belongs to $T_3$ class of estimators. The famous Horvitz-Thompson estimator of population total, $e_s = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$ also belongs to $T_2$ class of estimators.

**Use of auxiliary information**

Many a time in the survey additional information is available on another variable, related with the study variable. This variable is known as an auxiliary variable. Generally, the auxiliary variable is highly correlated with the study variable. It is also assumed that the information on the auxiliary variable is complete i.e. the information on auxiliary variable is available for all the population units. Further, the information on the auxiliary variable is available before the sample is selected. For instance, the study variable in a survey may be holding-wise irrigated area and it is proposed to estimate the total irrigated area in a district. The auxiliary variable in this case could be the holding-wise cultivated area in the district. A high correlation is expected between the variables irrigated area and cultivated area. This a-priori information on holding-wise cultivated area can be used at the estimation stage to develop a more precise estimator of total irrigated area in the district. Similarly, let the study variable be the number of tube wells in different villages and the objective is to estimate the total number of tube wells in a district. Both the variables village-wise number of tube wells observed in previous year or village-wise net irrigated area in previous year in the district are expected to be correlated with the village-wise number of tube wells (study variable) observed in current year. The two variables can be taken as auxiliary variables for the study variable to improve the precision of the estimator of total number of tube wells in a district in
the current year. Likewise, if the study variable is household-wise quantity of milk produced and the objective is to estimate quantity of milk produced in a large area, the variable household-wise number of cattle is expected to be correlated with the study variable and thus may be used as auxiliary variable for improving the precision of the estimator. Information on the auxiliary variable can be collected relatively inexpensively when not available and can be gainfully employed for improving the precision of the estimator.

The auxiliary information, if available, can be utilized in a many ways in the surveys. As described above in the section on unequal probability sampling, the auxiliary information can be used in the selection of the sample. Many a time, the sampling units are found to be of varying sizes. For example, in a sample survey conducted for estimation of crop production, the larger villages may have greater area under the crop and thus the contribution of larger villages to the total production is expected to be more. For such situations the precision of the estimator can be increased by selecting sampling units (villages) with unequal probabilities. The sampling units can be selected with probabilities directly proportional to their sizes, i.e.,

$$P_i \propto X_i \text{ or } P_i = \frac{X_i}{X_1 + \cdots + X_i + \cdots + X_N}.$$ 

Thus the available auxiliary information on size of the sampling units can be made use of in increasing the precision of the estimator.

The auxiliary information is also frequently used at the estimation stage to improve the precision of the estimators. Some examples of estimators of the population mean (or population total) that make use of auxiliary information at the estimation stage are ratio and regression estimators. These estimators are very precise when there is high correlation between the study and auxiliary variable. In what follows, we provide the relevant details on the ratio and regression estimators.

**Ratio estimator**

The ratio estimator is commonly used for estimation of population total $Y$ or population mean $\bar{Y}$.

Suppose that a SRSWOR of size $n$ is drawn from a population $U$ containing $N$ distinct and identifiable units. Let $(y_1, x_1), (y_2, x_2), \cdots, (y_i, x_i), \cdots, (y_n, x_n)$ denote the observations on the study variable and the auxiliary variable, respectively. The ratio estimator of population mean is given by
\[ \hat{Y}_r = \frac{\bar{Y}}{\bar{X}} \]; where \( \bar{Y} \) and \( \bar{X} \) are the sample means of \( y \) and \( x \), respectively.

while the estimator for population total is given by

\[ \hat{Y}_r = N \frac{\bar{Y}}{\bar{X}}. \]

Henceforth we shall restrict the description to the estimation of population mean.

The estimator \( \hat{Y}_r \) is biased, with expression for the bias to the first order of approximation given by

\[ \text{Bias}\left( \hat{Y}_r \right) = \left( \frac{1}{n} - \frac{1}{N} \right) \frac{\bar{Y}}{N} \left( \frac{C_x^2}{\rho} - C_y \right) \]

The mean square error \( \text{MSE}\left( \hat{Y}_r \right) \), of the ratio estimator \( \hat{Y}_r \) of population mean \( \bar{Y} \), to the first order of approximation is given by

\[ \text{MSE}\left( \hat{Y}_r \right) = \left( \frac{1}{n} - \frac{1}{N} \right) \left( s_x^2 + R_n^2 s_y^2 - 2R_n s_{xy} \right) \]

Here \( R_n = \frac{\bar{Y}}{\bar{X}} \); \( S_{xy} = (N - 1)^{-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y}) \),

and \( S_x^2 = (N - 1)^{-1} \sum_{i=1}^{N} (X_i - \bar{X})^2 \), \( u = x, y \).

An unbiased estimator of the mean square error, \( \text{MSE}\left( \hat{Y}_r \right) \), is given by

\[ \text{MSE}\left( \hat{Y}_r \right) = \left( \frac{1}{n} - \frac{1}{N} \right) \left( s_x^2 + R_n^2 s_y^2 - 2R_n s_{xy} \right) \]

Here \( R_n = \frac{\bar{Y}}{\bar{X}} \); \( s_{xy} = (n - 1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \),

and \( s_x^2 = (n - 1)^{-1} \sum_{i=1}^{n} (u_i - \bar{u}_n)^2 \), \( u = x, y \).

Similarly, the bias of the ratio estimator of population total is given by

\[ \text{Bias}\left( \hat{Y}_r \right) = N \times \text{Bias}\left( \hat{Y}_r \right) = N \left( \frac{1}{n} - \frac{1}{N} \right) \frac{\bar{Y}}{N} \left( \frac{C_x^2}{\rho} - C_y \right) \]
and $MSE(\hat{Y}_R)$ to the first order of approximation is given by

$$MSE(\hat{Y}_R) = N^2N\widehat{MSE}(\hat{Y}_R) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)(s_x^2 + R_{X,Y}^2s_y^2 - 2R_{X,Y}s_Xs_Y).$$

An unbiased estimator of the $MSE(\hat{Y}_R)$ is given by

$$M\bar{S}E(\hat{Y}_R) = N^2\bar{M}SE(\hat{Y}_R) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)(\bar{s}_y^2 + R_{X,Y}^2s_x^2 - 2R_{X,Y}s_Xs_Y).$$

The ratio estimator of population mean (population total) is more efficient than the corresponding estimator under SRSWOR sampling design [an estimator of population mean (population total) which does not take into account the auxiliary information] if

$$\rho > \frac{1}{2} \frac{C_x}{C_y}.$$  \hspace{1cm} (29)

It may be seen that for larger values of correlation coefficient between $Y$ and $X$, the condition in (29) will hold easily. Thus, ratio estimator is likely to be more precise if the correlation coefficient between $Y$ and $X$ is high.

**The regression estimator**

It has been just described that the ratio method of estimation performs well when there is a linear relationship between the study variable and the auxiliary variable and the correlation coefficient between the two variables is positive and large. However, no mention was made about the intercept of the linear relationship. If the intercept is large, then the precision of estimation can be further improved by making use of regression estimator.

Suppose that a SRSWOR of size $n$ is drawn from a population $U$ containing $N$ distinct and identifiable units. Let $(y_1,x_1),(y_2,x_2),\ldots,(y_i,x_i),\ldots,(y_n,x_n)$ denote the observations on the study variable and the auxiliary variable, respectively. The regression estimator of population mean, $\bar{Y}_{lr}$, is given by

$$\hat{Y}_{lr} = \bar{y}_n + \hat{\beta}(\bar{x}_n - \bar{x}_n),$$

where $\hat{\beta}$ is the estimator of the population regression coefficient $\beta$ and is given by
The regression estimator is biased and the bias is given by

\[ Bias(\hat{Y}_{lr}) = -\text{Cov}(\hat{\beta}, \bar{Y}_n). \]

The bias vanishes if the joint distribution of \( X \) and \( Y \) is bivariate normal.

The approximate expression for the mean square error of the regression estimator is given by

\[ \text{MSE}(\hat{Y}_{lr}) = \left(1 - \rho^2\right)\left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 < \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2, \]

so long as \( \rho \neq 0 \). Here \( \rho = \frac{S_{xy}}{S_xS_y} \).

An estimator of the mean square error of the regression estimator is given by

\[ \text{MSE}(\hat{Y}_{lr}) = (1-r^2)\left(\frac{1}{n} - \frac{1}{N}\right)S_y^2, \]

where \( r = \frac{s_{xy}}{S_xS_y} \).

It is easy to see that the mean square error of regression estimator is smaller than that of ratio estimator and is also smaller that the variance of the sample mean under SRSWOR sampling design.

**Regression estimator of population total**

The expressions relating to estimators of population total can be easily obtained from the expressions for the estimator of population mean given above and can be written as follows:

\[ \hat{Y}_{lr} = N\hat{\bar{Y}}_{lr} = N\left[\bar{y}_n + \hat{\beta}(\bar{X} - \bar{x}_n)\right] \]

\[ Bias(\hat{Y}_{lr}) = N\text{Bias}(\hat{\bar{Y}}_{lr}) = -NCov(\hat{\beta}, \bar{Y}_n). \]

\[ \text{MSE}(\hat{Y}_{lr}) = N^2 \times \text{MSE}(\hat{\bar{Y}}_{lr}) = N^2\left(1 - \rho^2\right)\left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 \]

\[ \text{MSE}(\hat{Y}_{lr}) = N^2 \times \text{MSE}(\hat{\bar{Y}}_{lr}) = N^2\left(1 - r^2\right)\left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 \]
Auxiliary information can also be used for better planning of surveys. Thus, if the targeted population is heterogeneous, it is possible to divide the population into number of homogeneous groups called as strata on the basis of available auxiliary information and samples can be selected from each of the stratum. The ultimate sample selected from different groups (strata) is expected to be a better representative of the population and is expected to be more precise compared to SRSWOR.

**Stratified sampling**

In case of simple random sampling without replacement, the sampling variance of the sample mean is

\[
V(\bar{y}_n) = \left(1 - \frac{1}{N}\right)S^2_y.
\]

Clearly, the variance decreases as the sample size \((n)\) increases or the population variability \(S^2\) decreases. However, a good sampling strategy is one which helps in reducing the sampling variance to the lowest possible extent. As a matter of fact \(S^2_y\) is a population parameter (in fact it is a measure of population variance) and is a constant quantity. The other way of increasing the precision of estimation is to increase the sample size. The sample size also cannot be increased because it would result in an increase in the cost of the survey. An increase in sample size may lead to other type of errors, like non-sampling errors. Instead, a way to increase the precision of estimation is through a better representativeness of the population in the sample. The population may be divided into several homogeneous groups called strata and then sampling may be carried out independently within each stratum. This way the within stratum variability is reduced and the between strata variability is taken care of by independent sampling within each stratum. This procedure of sample selection is known as stratified sampling.

Stratified sampling is a very popular procedure in sample surveys. The procedure enables one to draw a sample with any desired degree of representation of the different parts of the population by taking them as strata. In stratified sampling, the population consisting of \(N\) units is first divided into \(K\) disjoint sub-populations of \(N_1, N_2, \ldots, N_K\) units, respectively. These sub-populations are non-overlapping and together they comprise the whole of the population.
i.e. \( \sum_{i=1}^{K} N_i = N \). These sub-populations are called strata. To obtain full benefit from stratification, the values of \( N_i \)'s must be known. When the strata have been determined, a sample is drawn from each stratum, the drawings being made independently in different strata. If a simple random sample without replacement is taken from each stratum, then the procedure is termed as stratified random sampling. Since the sampling is done independently from each stratum, the precision of the estimator of the population mean or the population total from each stratum would depend upon the variability within each stratum. Thus the stratification of population should be done in such a way that the within stratum variability is as small as possible and the between strata variability is as large as possible. Thus, the strata are homogeneous within themselves with respect to the variable under study. However, in many practical situations it is usually difficult to stratify with respect to the variable under consideration especially because of geographical and cost considerations. Generally, the stratification is done according to administrative groupings, geographical regions and on the basis of auxiliary characters correlated with the character under study.

Suppose that a SRSWOR sample of size \( n \) is drawn from a population containing \( N \) units stratified into \( K \) strata of respective sizes \( N_1, N_2, \ldots, N_K \), such that \( \sum_{i=1}^{K} N_i = N \). Let \( y_{ij} \) denote the \( j^{th} \) observation in the \( i^{th} \) stratum.

The population mean can be expressed as

\[
\bar{Y} = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{N_i} y_{ij}.
\]

Again the population mean square is \( S^2 = (N-1)^{-1} \sum_{i=1}^{K} \sum_{j=1}^{N_i} \left( y_{ij} - \bar{Y} \right)^2 \).

We define population mean square for the \( i^{th} \) stratum as \( S_i^2 = (N_i - 1)^{-1} \sum_{j=1}^{N_i} \left( y_{ij} - \bar{Y}_{N_i} \right)^2 \),

where \( \bar{Y}_{N_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij} \) is the population mean for the \( i^{th} \) stratum.

Draw SRSWOR of sizes \( n_1, n_2, \ldots, n_i, \cdots, n_k \), respectively, such that \( n_1 + n_2 + \cdots + n_i + \cdots + n_k = n \) (sample size). 

23
The population mean $\bar{Y}$ can be rewritten as

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{K} N_i \bar{y}_i = \sum_{i=1}^{K} P_i \bar{Y}_i$$

where $P_i = N_i/N$. Since the sampling has been carried out independently within each stratum by SRSWOR, sample mean for the $i$th stratum, $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$, is an unbiased estimator of $\bar{Y}_i$ and obviously the stratified sample mean

$$\bar{y}_st = \frac{1}{N} \sum_{i=1}^{K} N_i \bar{y}_i = \sum_{i=1}^{K} P_i \bar{y}_i,$$

which is the weighted mean of the strata sample means with strata size as the weights, will be an appropriate estimator of the population mean $\bar{Y}$.

Thus, $\bar{y}_st$ is an unbiased estimator of $\bar{Y}$.

Since the sample in the $i$th stratum has been drawn by SRSWOR, so

$$V(\bar{y}_{ni}) = \left( \frac{1}{n_i} - \frac{1}{N_i} \right) s_i^2.$$  

The sampling variance of $\bar{y}_st$ is given by

$$V(\bar{y}_st) = \sum_{i=1}^{K} P_i^2 V(\bar{y}_i) = \sum_{i=1}^{K} P_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) s_i^2. \quad (30)$$

Since the sample mean square for the $i$th stratum, $s_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{ni})^2$ is an unbiased estimator of $S_i^2$, it follows that an unbiased estimator of $V(\bar{y}_st)$ is given by

$$\hat{V}(\bar{y}_{st}) = \sum_{i=1}^{K} P_i^2 \hat{V}(\bar{y}_i) = \sum_{i=1}^{K} P_i^2 \left( \frac{1}{n_i} - \frac{1}{N_i} \right) s_i^2.$$ 

From the above expression it is clear that the sampling variance of stratified sample mean depends on $S_i^2$, the variability within the strata. It, therefore, follows that homogeneous strata will lead to a greater precision of the stratified sample mean.
Allocation of sample size to different strata

In stratified sampling, having decided upon the number of strata, the strata boundaries and the total sample size \( n \) to be drawn, the next question which a survey statistician has to answer is regarding the allocation of total sample of size \( n \) to respective strata and also the method of selection of the allocated sample within each stratum. The expression for the variance of stratified sample mean shows that the precision of a stratified sample for given strata depends upon the \( n_i \)'s, which can be fixed at will. The guiding principle in the determination of the \( n_i \)'s is to choose them in such a manner so as to provide an estimate of the population mean with the desired degree of precision for a minimum cost or to provide an estimate with maximum precision for a given cost, thus making the most effective use of the available resources. The allocation of the sample to different strata made according to this principle is called the principle of optimum allocation.

The cost of a survey is a function of strata sample sizes, \( n_i \)'s, just as the variance is. However, the purpose of the survey and the nature of the study variable will dictate the manner in which the cost of the survey will vary with total sample size. Similarly, the allocation of the sample size to different strata will also depend upon the purpose of the survey and the nature of the study variable. In yield estimation surveys, the major item in the survey cost consists of labour charges for harvesting of produce and as such survey cost is found to be approximately proportional to the number of crop cutting experiments (CCE). The cost per CCE may, however, vary in different strata depending upon labour availability. Under such situations, the total cost may be represented by

\[
C = \sum_{i=1}^{K} c_i n_i,
\]

where, \( c_i \) is the total cost of a CCE in the \( i^{th} \) stratum. When \( c_i \) is same from stratum to stratum, i.e., \( c_i = c \ \forall \ i = 1,2,\cdots,K \), then the total cost of a survey is given by

\[
C = cn.
\]

The cost function will change in form if the travel cost, salary of the field staff, the cost involved in statistical analysis of data, etc. are to be paid for.

The optimum values of \( n_i \)'s can be obtained by minimizing the \( V(\bar{y}_n) \) for fixed cost \( C \) as
\[ \frac{P_i S_i}{\sqrt{n_i}} = \sqrt{\mu c_i n_i}, \text{ or } n_i = \frac{P_i S_i}{\sqrt{\mu c_i}}, i = 1, 2, \cdots, K, \]

where \( \mu \) is some constant.

From the above, one can easily infer that:

- the larger the stratum size, the larger should be the size of the sample to be selected from that stratum;
- the larger the stratum variability, the larger should be the size of the sample from that stratum and
- the cheaper the cost per sampling unit in a stratum, the larger should be the sample from that stratum.

The exact value of \( n_i \), obtained after evaluating \( \frac{1}{\sqrt{\mu}} \), the constant of proportionality for maximizing precision for a fixed cost \( C_0 \), is given by

\[ n_i = \frac{P_i S_i}{\sqrt{c_i}} \frac{C_0}{\sum_{i=1}^{K} P_i S_i \sqrt{c_i}}, i = 1, 2, \cdots, K \]  \hspace{1cm} (31)

where \( C_0 \) is the total cost of the survey. The total sample size, \( n \), is given by

\[ n = \sum_{i=1}^{K} n_i = \frac{C_0 \sum_{i=1}^{K} (P_i S_i / \sqrt{c_i})}{\sum_{i=1}^{K} P_i S_i \sqrt{c_i}} \]  \hspace{1cm} (32)

The allocation of sample size \( n \) according to above equation is known as optimum allocation.

When \( c_i \) is same from stratum to stratum, i.e., \( c_i = c \ \forall \ i = 1, 2, \cdots, K \), the cost function takes the form \( C = cn \), or in other words, the cost of survey is proportional to the size of the sample. The optimum values of \( n_i \)'s are then given by

\[ n_i = n \frac{P_i S_i}{\sum_{i=1}^{K} P_i S_i}, i = 1, 2, \cdots, K. \]  \hspace{1cm} (33)

The allocation of the sample according to the above formula is known as Neyman Allocation.

Using \( n_i \) from (33) in the \( V(\bar{Y}_u) \) expression in (30), gives
where the subscript \(NY\) stands for the stratification with Neyman Allocation.

The stratum variability, \(S_i\), being a population parameter is generally not known. So another approach of sample allocation is

\[ n_i \alpha N_i \text{ or } n_i = n \frac{N_i}{N} \quad i = 1, 2, \ldots, K, \tag{34} \]

\(i.e.,\) to allocate larger sample sizes for larger strata. The allocation of sample size \(n\) according to (34) is known as proportional allocation and \(V(\bar{y}_{ny})\) in this case becomes

\[ V_p(\bar{y}_{ny}) = \left( \frac{1}{n} - \frac{1}{N} \right) \sum_{i=1}^{K} P_i S_i^2, \]

where the subscript \(P\) indicates the stratification with proportional allocation.

**Cluster sampling**

The smallest units into which the population can be divided are called the elements of the population and groups of these units/elements are called clusters. A cluster may be a class of students or cultivators fields in a village. When sampling unit is a cluster, the procedure of sampling is called cluster sampling. Cluster sampling is a sampling technique in which the entire population of interest is divided into clusters, and a sample of these clusters is selected by SRSWOR technique.

The main reason for using cluster sampling is that it is usually much cheaper and more convenient to sample the clusters rather than individual units. Many a times, constructing a sampling frame that identifies every population element is too expensive or impossible to construct. For example, list of all farms in a district is generally not available but the list of villages may be easily available. Considering villages as clusters, selection of villages can be done by SRSWOR and then complete enumeration of the selected villages can be done. Cluster sampling can also reduce the cost of survey when the population elements are scattered over a wide area.

For a given number of sampling units, cluster sampling is more convenient and less costly than element sampling due to the saving time in journeys, identification and contacts, etc. Cluster sampling, however, is generally less efficient than SRSWOR sampling design with
sample mean as the estimator of the population mean due to the tendency of the units in a cluster to be similar. In most practical situations, the loss in efficiency may be more than offset by the reduction in the cost and the efficiency per unit cost may be more in cluster sampling as compared to element sampling.

The basic task in cluster sampling is to specify appropriate clusters. Clusters are generally made up of neighboring units or of compact areas and, therefore, the units within the clusters tend to have similar characteristics. As a simple rule, the number of elements in a cluster should be small and the number of clusters should be large. The efficiency of cluster sampling decreases with the increase in the size of the cluster. Selection of required number of clusters may be done by equal or unequal probabilities of selection and all selected clusters are required to be completely enumerated. The selection of the clusters can also be made by first selecting randomly a unit, called the key unit, and then taking randomly the required number of neighboring clusters of the key clusters as the sample of clusters. For example for estimating the milk production, cluster of three villages may be formed by first selecting a key village at random and then taking two more villages amongst the villages in a circle of some specified radius with the key village as the center village.

Notations

As before, let $N$ denote the number of clusters in the population. Suppose that a SRSWOR of $n$ clusters is to be selected in the sample. Let $M_i$ denote the number of elements in the $i^{th}$ cluster, $i = 1, 2, \cdots, N$. We shall also denote by $M_o = \sum_{i=1}^{N} M_i$, the total number of elements in the population and by $\bar{M} = M_o/N$ the average number of elements per cluster in the population. Let $y_{ij}$ denote the value of the character under study for the $j^{th}$ element in the $i^{th}$ cluster $j = 1, 2, \cdots, M_i; i = 1, 2, \cdots, N$). The population mean per element of the $i^{th}$ cluster is then given by $\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$, $i = 1, 2, \cdots, N$. Then the population mean of the cluster means is given by $\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_i$. The population mean per element is $\bar{Y} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}}{M_o}$.
**Equal cluster sizes**

For simplicity, the case of equal cluster sizes, *i.e.*, \( M_i = M \) for all \( i = 1, 2, \ldots, N \) may be considered first. When the selection is by SRSWOR, the unbiased estimator of the population mean \( \bar{Y} \) for equal cluster sizes is given by

\[
\bar{y}_{cl} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i.
\]

The variance of \( \bar{y}_{cl} \) is given by

\[
V(\bar{y}_{cl}) = \left( \frac{1}{n} - \frac{1}{N} \right) S_b^2,
\]

where \( S_b^2 = (N-1)^{-1} \sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2 \).

An unbiased estimator of variance of \( \bar{y}_{cl} \) is given by

\[
\hat{V}(\bar{y}_{cl}) = \left( \frac{1}{n} - \frac{1}{N} \right) s_b^2
\]

where \( s_b^2 = (n-1)^{-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{cl})^2 \).

For the \( n \) selected clusters, the estimator of population mean, \( \bar{y}_{cl} \), is based on a sample of \( nM \) elements. Hence the relative efficiency of \( \bar{y}_{cl} \) with respect to \( \bar{y} \), the sample mean based on a SRSWOR sample of size \( nM \) elements selected from a population of \( NM \) elements is

\[
RE(\bar{y}_{cl}) = \frac{\left( \frac{1}{nM} - \frac{1}{NM} \right) S^2}{\left( \frac{1}{n} - \frac{1}{N} \right) S_b^2} = \frac{S^2}{MS_b^2},
\]

where \( S^2 = (NM - 1)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2 \).

Thus it can be seen that the relative efficiency of the cluster sampling increases as the cluster size decreases. Similarly, the efficiency increases as overall mean square between elements increases and mean square between clusters decreases.
Unequal cluster sizes

Clusters are generally of unequal sizes. In what follows, we provide the theoretical details of this setting.

When clusters are of unequal sizes, an estimator of population mean $\bar{Y}$ can be taken as

$$\bar{y}_{cl}^* = \frac{1}{n} \sum_{i=1}^{n} \bar{Y}_i$$

which is an unbiased estimator for $\bar{Y}_N$ and not for $\bar{Y}$. The expression for variance (MSE) and variance estimator for above estimator are given by

$$V\left(\bar{y}_{cl}^*\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{b2}^*,$$

where $S_{b2}^* = (N - 1)^{-1} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y}_N)^2$,

and $\hat{V}\left(\bar{y}_{cl}^*\right) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{b2}^*$

where $s_{b2}^* = (n - 1)^{-1} \sum_{i=1}^{n} (\bar{Y}_i - \bar{y}_{cl}^*)^2$.

In the present situation, an unbiased estimator of population mean $\bar{Y}$ may be obtained by

$$\bar{y}_{cl} = \frac{1}{n} \sum_{i=1}^{n} \frac{M_i}{M} \bar{Y}_i.$$  

The variance and the estimator of the variance of $\bar{y}_{cl}$ are obtained as

$$V\left(\bar{y}_{cl}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{b2}^{'},$$

where $S_{b2}^{' = (N - 1)^{-1} \sum_{i=1}^{N} \left(\frac{M_i}{M} \bar{Y}_i - \bar{Y}\right)^2},$

and $\hat{V}\left(\bar{y}_{cl}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{b2}^{'}$.
where $s_h^2 = (n-1)^{-1} \sum_{i=1}^{n} \left( \frac{M_i \bar{y}_i - \bar{y}_{cl}}{M} \right)^2$.

**Multi-stage sampling**

Cluster sampling is a sampling procedure in which clusters are considered as sampling units and all the elements of the selected clusters are enumerated. One of the main considerations of adopting cluster sampling is the reduction of travel cost because of the nearness of elements in the clusters. However, this method restricts the spread of the sample over population which results generally in increasing the variance of the estimator. In order to increase the efficiency of the estimator with the given cost it is natural to think of further sampling the clusters and selecting more number of clusters so as to increase the spread of the sample over population. This type of sampling which consists of first selecting clusters and then selecting a specified number of elements from each selected cluster is known as subsampling or two stage sampling, since the units are selected in two stages. In such sampling designs, clusters are generally termed as first stage units (fsu’s) or primary stage units (psu’s) and the elements within clusters or ultimate observational units are termed as second stage units (ssu’s) or ultimate stage units (usu’s). It may be noted that this procedure can be easily generalized to give rise to multistage sampling design, where the sampling units at each stage are clusters of units of the next stage and the ultimate observational units are selected in stages, sampling at each stage being done from each of the sampling units or clusters selected in the previous stage. This procedure, being a compromise between uni-stage or direct sampling of units and cluster sampling, can be expected to be (i) more efficient than uni-stage sampling and less efficient than cluster sampling from considerations of operational convenience and cost, and (ii) less efficient than uni-stage sampling and more efficient than cluster sampling from the view point of sampling variability, when the sample size in terms of number of ultimate units is fixed.

It may be mentioned that multistage sampling may be the only feasible procedure in a number of practical situations, where a satisfactory sampling frame of ultimate observational units is not readily available and the cost of obtaining such a frame is prohibitive or where the cost of locating and physically identifying the usu’s is considerable. For instance, for conducting a socio-economic survey in a region, where generally household is taken as the usu, a complete and up-to-date list of all the households in the region may not be available,
whereas a list of villages and urban blocks, which are group of households, may be readily available. In such a case, a sample of villages or urban blocks may be selected first and then a sample of households may be drawn from each selected village and urban block after making a complete list of households. It may happen that even a list of villages is not available, but only a list of all tehsils (group of villages) is available. In this case a sample of households may be selected in three stages by selecting first a sample of tehsils, then a sample of villages from each selected tehsil after making a list of all the villages in the tehsil and finally a sample of households from each selected village after listing all the households in it. Since the selection is done in three stages, this procedure is termed as three stage sampling. Here, tehsils are taken as first stage units (fsu’s), villages as second stage units (ssu’s) and households as third or ultimate stage units (tsu’s).

One of the advantages of this type of sampling is that at the first stage the frame of fsu’s is required which is generally easily available and at the second stage the frame of ssu’s is required for the selected fsu’s only, and so on. Moreover, this method allows the use of different selection procedures in different stages. It is because of these considerations that multistage sampling is used in most of the large scale surveys. It has been found to be very useful in practice. It is noteworthy that Prof. P. C. Mahalanobis used this sampling procedure in crop surveys carried out in Bengal during the period 1937-1941, and he had termed this procedure as nested sampling. Cochran (1939) and Hansen and Hurwitz (1943) have considered the use of this procedure in agricultural and population surveys respectively. Lahiri (1954) discussed the use of multistage sampling in the Indian Sample Survey.

Two stage sampling with equal probabilities, equal first stage units

Estimation procedure

Let the population under study consists of \(NM\) elements grouped into \(N\) first stage units, each first stage unit containing \(M\) second stage units.

Let us denote by \(y_{ij}\) the value of the characteristic under study for the \(j^{th}\) second stage unit of the \(i^{th}\) first stage unit, \(j=1,2,\cdots,M; i=1,2,\cdots,N\). Let the mean of the \(i^{th}\) fsu be denoted by

\[
\bar{Y}_i = \frac{1}{M} \sum_{j=1}^{M} y_{ij},
\]

and the population mean by

\[
\bar{Y} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_i.
\]
Further, let a sample of size $nm$ be selected by first selecting $n$ fsu’s from $N$ fsu’s by SRSWOR and then selecting $m$ ssu’s from $M$ ssu’s by SRSWOR from each of the selected fsu’s. Let us denote by $\bar{y}_{im} = \frac{1}{m} \sum_{j=1}^{m} y_{ij}$ the sample mean based on $m$ selected ssu’s from the $i^{th}$ fsu, and by $\bar{y}_{nm} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_{im}$, the sample mean based on all the $nm$ units in the sample.

Clearly, $\bar{y}_{nm}$ is an unbiased estimator of $\bar{y}$ with its variance given by

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) s^2_b + \frac{1}{n} \left(\frac{1}{m} - \frac{1}{M}\right) s^2_w$$

where $s^2_b = (N-1)^{-1} \sum_{i=1}^{N} (\bar{y}_i - \bar{y})^2$

and $s^2_w = \frac{1}{N} \sum_{i=1}^{N} S^2_i = \frac{1}{N(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_i)^2$

The estimator of $V(\bar{y}_{nm})$ is given by

$$\hat{V}(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) s^2_b + \frac{1}{n} \left(\frac{1}{m} - \frac{1}{M}\right) s^2_w$$

where $s^2_b = (n-1)^{-1} \sum_{i=1}^{n} (\bar{y}_{im} - \bar{y}_{nm})^2$

and $s^2_w = \frac{1}{n} \sum_{i=1}^{n} S^2_i = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \bar{y}_{im})^2$.

It is observed that the variance of sample mean ($\bar{y}_{nm}$) in two stage sampling consists of two components; the first representing the contribution arising from sampling of first stage units and the second arising from sub-sampling within the selected first stage units. We note the following two cases:

Case (i) $n = N$, corresponds to stratified sampling with $N$ first stage units as strata and $m$ units drawn from each stratum.
Case (ii) \( m = M \), corresponds to cluster sampling.

**Two stage sampling unequal first stage units**

**Estimation procedure**

The situation of two-stage sampling where primary stage units have varying number of second-stage units is most likely to be encountered in practical situations. Thus, villages in a district may have varying number of households/cultivator fields. Let the population under consideration consist of \( N \) primary sampling units and \( i \)th primary sampling unit contains \( M_i \) secondary sampling units. Further, suppose that a sample of \( n \) fsu’s is selected from \( N \) fsu’s by SRSWOR and from the \( i \)th selected fsu, a sample of \( m_i \) ssu’s is selected from \( M_i \) ssu’s by SRSWOR.

Let us denote by \( y_{ij} \) the value of character under study for the \( j \)th ssu of the \( i \)th fsu of the population, \( j = 1,2,\cdots,M_i; i = 1,2,\cdots,N \). Let the population mean of the \( i \)th fsu be denoted by

\[
\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij},
\]

and the population mean by \( \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} M_i \bar{Y}_i \). Further, let \( \bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i \) denote the average number of ssu’s per fsu and \( \bar{y}_{im} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \), the sample mean for i-th fsu.

An unbiased estimator of the population mean, \( \bar{Y} \), in the present situation is given by

\[
\bar{y}_s = \frac{1}{n \bar{M}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij}.
\]

The above estimator is useful in situations where information on total number of second-stage units within the selected first stage units is available i.e. \( M_i \) is known for \( i = 1,2,\cdots,n \).

The variance of \( \bar{y}_s \) is given by

\[
V(\bar{y}_s) = \left( \frac{1}{n} - \frac{1}{N} \right) S_y^2 + \frac{1}{nN} \sum_{i=1}^{N} \frac{M_i^2}{\bar{M}^2} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_i^2.
\]
where \( S_b^2 = (n-1)^{-1} \sum_{i=1}^{N} \left( \frac{M_i}{M} \bar{y}_i - \bar{y}_{ts} \right)^2 \)

and \( S_i^2 = \frac{1}{(M_i-1)} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_i)^2 \).

The estimator of \( V(\bar{y}_{ts}) \) is given by

\[
\hat{V}(\bar{y}_{ts}) = \left( \frac{1}{n} - \frac{1}{N} \right) S_b^2 + \frac{1}{nN} \sum_{i=1}^{N} \frac{M_i^2}{M^2} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_i^2
\]

where \( S_b^2 = (n-1)^{-1} \sum_{i=1}^{N} \left( \bar{y}_i - \bar{y}_{mm} \right)^2 \), \( \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \),

and \( S_i^2 = \frac{1}{m_i-1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{im})^2 \).

Another estimator of the population mean which is biased is given by

\( \bar{y}_{**} = \frac{1}{n} \sum_{i=1}^{N} \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \)

It is important to note that the above estimator does not require availability of information on the total number of second-stage units in the selected first-stage units. Since the above estimator is biased, relevant criterion for judging the suitability of the estimator is the mean square error.

The mean square error of \( \bar{y}_{ts} \) is given by

\[
MSE(\bar{y}_{ts}) = \left( \frac{1}{n} - \frac{1}{N} \right) S_b^2 + \frac{1}{nN} \sum_{i=1}^{N} \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_i^2,
\]

where \( S_b^2 = (N-1)^{-1} \sum_{i=1}^{N} (\bar{y}_i - \bar{y})^2 \).

The estimator of \( MSE(\bar{y}_{**}) \) is given by
\[ MSE(\bar{y}_{ts}) = \left( \frac{1}{n} - \frac{1}{N} \right) s_b^2 + \frac{1}{nN} \sum_{i=1}^{n} \left( \frac{1}{m_i} - \frac{1}{M} \right) s_i^2 \]

where \( s_b^2 = (n-1)^{-1} \sum_{i=1}^{n} \left( \bar{y}_i - \bar{y}_{ts} \right) \) and \( \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \).

When \( M_i = M \) for all \( i \), both the estimators considered above are the same.

**References**


