Chapter 2

Preliminaries

Summary

We will begin this chapter with some basic concepts and definitions which will be used in the chapters to follow. The superpopulation model concept is next introduced in brief. The utilisation of auxiliary information is described mainly at stages - for selection of sample as size measure in PPS sampling and for ratio method of estimation in stratified sampling.

2.1 Concepts and Definitions

In this section, we give some concepts and definitions which will be used in this thesis. By finite population $U$, we mean a collection of units $U_1, U_2, \ldots, U_N$, where $N$ is a known finite number and the units $U_1, U_2, \ldots, U_N$, are identifiable or distinguishable.
We denote the finite population by

$$U = (U_1, U_2, \ldots, U_N)$$

(2.1.1)
a list of such units as in (2.1.1) is called a "sampling frame" and N is called
the "size of the population".

A sample space

$$S = \{s\}$$

(2.1.2)
is the collection of all possible samples s from U where either $S = \{s\}$ is
the collection of all non-empty subsets s of units from U in which case s is a
non-empty subset of U i.e.,

$$s = (U_1, U_2, \ldots, U_{n(s)})$$

(2.1.3)
or

$S = \{s\}$ is the collection of all finite ordered sequences s of units from U in
which case

$$s = (U_{i_1}, U_{i_2}, \ldots, U_{i_{n(s)})}$$

(2.1.4)
the number of units in s i.e., the cardinality of s, denoted by $n(s)$ is called
the sample size of s. Thus, if s is a subset of U, $n(s)$ is the number of units
in s and s is a sequence of units from U, $n(s)$ denotes the length of the
sequence s.

Consider a real valued variable Y (study variable) defined over $U$ and
taking value $Y_i$ on unit $i$, $1 \leq i \leq N$. Let $\mathbf{Y}$ denote the vector

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$$  \hspace{1cm} (2.1.5)

the $Y_i$'s are unknown apriori and general problem in sample survey is to estimate real valued functions of $\mathbf{Y}$ the parameters of interest called parametric functions on the basis of the observations $y_i$ for $i \in s$ where $s$ is a sample drawn with given probability $p_s$ from the sample space $S$, the totality of all possible samples $s$. Let $P$ be a probability measure defined on $S$ such that

$$p_s \geq 0, \text{ for all } s \in S \text{ and } \sum_{s \in S} p_s = 1$$  \hspace{1cm} (2.1.6)

where $p_s$ is the probability of selecting sample $s$.

The pair $(S, P)$ is called “Sampling Design” and is denoted by

$$D = D(S, P) = (S, P).$$  \hspace{1cm} (2.1.7)

The sampling design $D(S, P)$ is also denoted by $p$.

The definition of sampling design provides us a method of selecting a sample for which it is necessary to construct sample space of all possible samples and to select one of them with the corresponding probabilities prescribed by a sampling design $p$. However, it is not possible to put this method in practice, especially in large scale surveys. But Hanurav (1962) [22] demonstrated that every sampling design $D(S, P)$ can be implemented
by some practically convenient method called "Sampling Scheme". A sampling scheme facilitates drawing of units from $U$ one by one at random with probabilities which results in selection of sample $s$ from $S$ with desired probabilities.

For a given design $p$ the first and second order "inclusion probabilities" are given by

$$
\pi_i(p) = \pi_i = \sum_{s \ni i} p_s \quad 1 \leq i \leq N \quad (2.1.8)
$$

and

$$
\pi_{ij}(p) = \pi_{ij} = \sum_{s \ni (i,j)} p_s \quad 1 \leq i \leq N \quad (2.1.9)
$$

where in (2.1.8) the summation in the right hand side is over all samples containing unit $i$ and in (2.1.9) the summation is over all samples that contain the units $i$ and $j$.

Considering the problem of estimation, and real valued function $t$ defined over a design $D = (S, P)$ such that for samples $s \in S$, the function $t_s(Y)$ depends only on the values of $Y$ for the units belonging to the sample is called a "statistic". A statistic $t_s(Y)$ or simply denoted by $t$ when used to estimate a parametric function $T(Y)$ is called an "estimator" of $T$.

We will denote $t_s(Y)$ by $t_s$ for $s \in S$ or simply $t$ if there is no scope for confusion. An estimator $t$ of $T$ is called an "unbiased estimator" if
\[ E(p : t) = \sum_{s \in S} t \ p_s = T \text{ for all } Y. \]  
(2.1.10)

When \( t \) is an unbiased estimator of \( T \), we say that \( t \) is unbiased for \( T \) or \( t \) is \( p \)-unbiased for \( T \) (i.e., \( t \) is unbiased for \( T \) with respect to \( p \)) an estimator which is not unbiased for \( T \) is called a “biased estimator” and its bias is given by

\[ B(t) = E(t) - T. \]  
(2.1.11)

In the estimation of \( T \), the deviation \( (t_s - T) \) is taken as “error” in the estimate \( t_s \) based on a sample \( s \). Any convex function \( f(t_s - T) \) is called “loss function” and \( E(f) \) is called the “expected loss”. An oft-used “loss function” is the “Mean Square Error” (MSE) given by

\[ M(p : t) = E(t - T)^2 = \sum_{s \in S} (t_s - T)^2 p_s \]  
(2.1.12)

where \( t_s \) is used as an estimator of \( T \) based on the sample \( s \) and \( p_s \) is the probability of selecting the sample \( s \). When \( t \) is unbiased estimator of \( T \), then \( M(p : t) \) is called “Sampling Variance” or simply “variance” of \( t \) which is the same as

\[ V(p : t) = \sum_{s \in S} t_s^2 p_s - T^2. \]  
(2.1.13)

For simplicity of notation, \( E(t) \), \( M(t) \), \( V(t) \) are also used to denote \( E(p : t) \), \( M(p : t) \), \( V(p : t) \) respectively when there is no ambiguity regarding design \( p \) from the context.
A design $D(S,P)$ together with an estimator $t$ of a parametric function $T$ defined over $D$ is called a “sampling strategy” (or a “strategy”) for the estimation of $T$ and is denoted by

$$H = H(D, t) = H(S, P, t) = H(p, t) = H(p : t).$$

(2.1.14)

This definition is due to Hajek (1958) but the importance of this terminology is stressed by Hanurav (1965,66) [23], [24].

A strategy $H(p : t)$ when used for the estimation of $T$ is called an “unbiased strategy” for $T$ if $t$ is $p$-unbiased for $T$. Otherwise, it is called a “Biased Strategy” for estimating $T$.

The expectation $E(H)$, variance $V(H)$ or mean square error $M(H)$ of a strategy for estimating $T$ are defined as the expectation, variance or mean square error of the corresponding estimator $t$ over the design $p$.

Given a design $D = D(S,P)$ and unbiased estimators $t_1$ and $t_2$ of the same parametric function $T$, both defined over $D$, $t_1$ is said to be “uniformly better” than $t_2$ if

$$M(p : t_1) \leq M(p : t_2) \text{ for all } Y$$

(2.1.15)

with strict inequality holding at least for one $Y$.

Also in a class $C$ of estimators of $T$ defined over a given design $D(S,P)$,
an estimator $t_1 \in C$ is the "best" estimator if $t_1$ is uniformly better than $t_2$ for all $t_2$ different from $t_1$ and belonging to $C$.

Thus we say that a strategy $H_1$ is better than another strategy $H_2$ for estimating $T$, if

$$M(H_1) \leq M(H_2) \text{ for all } Y,$$

with strict inequality at least for one $Y$.

Given a class $C$ of strategies for estimating $T$, a member $H_1$ in the class $C$ is called the "best" if it is better than every other member of $C$ or in other words, a strategy $H_1$ is the "best" in class $C$ if

$$M(H_1) \leq M(H_2) \text{ for all } Y,$$

with strict inequality at least for one $Y$ and for all $H_2 \neq H_1$ and belonging to $C$.

Given a design $D = D(S,P)$, a "cost function" which is reasonable is given by

$$C_s = C_0 + C_1 \ n(s)$$

for every $s \in S$ where $C_s$ is the cost for a sample $s$, consisting of the overhead cost $C_0$, and the cost of collecting the data (the values of the study variable $Y$ on units in $s$) which is assumed to be proportional to the size of $s$ {i.e., $n(s)$} equal to $C_1 \ n(s)$ in which $C_1$ is the cost of collection of data on a single
The cost of a strategy $= (p : t)$ is the "expected cost of the design $p"$ and is given by

$$C(H) = C_0 + C_1 E\{n(s)\} \quad (2.1.19)$$

where

$$E\{n(s)\} = \sum_{s \in S} n(s) p_s \quad (2.1.20)$$

is the expected sample size of the design $D = D(S,P)$. It may be observed that under this set-up two strategies are "equally costly" if and only if, they have the same expected sample size. In this thesis we will consider only those strategies with $E\{n(s)\} = n$, a given number (i.e., strategies with expected sample size fixed).

### 2.2 Uses of Auxiliary Information

When an enquiry is planned through a sample survey, some auxiliary information is usually available. The auxiliary information can be utilized broadly at two stages viz., design stage and estimation stage in a sample survey. At the design stage, when the population units vary considerably in size, the auxiliary variable can be used for selection of sample by selecting units with unequal probabilities i.e., by "Probability Proportional to Size" (PPS) of the auxiliary variable. Another use of the auxiliary information can be made when the population under study is heterogenous w.r.t. the study variable. The population is divided with the help of auxiliary variable into
various homogeneous strata, referred to as the stratification. In the stratified sampling, there is another problem i.e., allocating the total sample over various strata termed as “allocation of sample size”. The problem of allocation of sample size can also be studied under the superpopulation model approach using the auxiliary information. At the estimation stage, the auxiliary information can be used by constructing various estimators such as ratio, product and regression estimators for improving the efficiency over the usual unbiased estimator.

2.3 Superpopulation Concept

An important step towards achieving an appropriate criterion of optimality is due to Cochran (1939, 1946) [4], [6]. Whenever information on an auxiliary real valued variable \( X \) taking positive value \( X_i \) on unit \( i, 1 \leq i \leq N \), which is highly correlated to the study variable \( Y \) is available, it is possible to use this information for setting up a criterion of optimality. We first introduce the concept of stochastic study variable let us pose a question: How are the population variable values generated? In an attempt to answer this question, we say that the \( N \) population values \( Y_1, Y_2, \ldots, Y_N \) of the study variable \( Y \) are generated from a superpopulation. The vector of population values \( \mathbf{Y} = (Y_1, Y_2, \ldots, Y_N) \) is assumed to be one particular realization of an \( N \)-length random vector \( \mathbf{Y} = Y_1, \ldots, Y_N \) of random variables. The joint distribution of \( \mathbf{Y} \), denoted by \( \delta \) is assumed to depend on \( X = (X_1, X_2, \ldots, X_N) \) and some unknown parameters. This concept is called Superpopulation Concept.
A "Superpopulation Model" or simply a "model" means a specified set of conditions that define a class of distributions to which \( \delta \) is assumed to belong. The finite population under study is treated to have been drawn from a larger universe of populations known as Superpopulation. This is the basic interpretation of idea of superpopulation in its most pure form. A model is the survey sampler's conceptualisation of superpopulation. In many situations, the model summarises and formalizes the model maker's prior knowledge about the study population whether it be based on long range experience or on personal subjective belief and is often an expression of prior knowledge or belief. The sampler is willing to consider the vector of finite population values \( \mathbf{Y} = (Y_1, Y_2, \ldots, Y_N) \) as a realisation of this superpopulation. Superpopulation modelling is the activity whereby one specifies or assumes certain features of the mechanism thought to have generated \( Y_1, Y_2, \ldots, Y_N \).

As mentioned earlier a "model" defines a class of populations or distributions specified to which \( \delta \), the joint distribution of \( \mathcal{Y} \) is assumed to belong. The specification may range from crude formulation, prescribing, for example only a few moments of the distribution \( \delta \) viz., the mean, variance and covariance of \( \delta \) to a highly detailed description of \( \delta \). We now present a few important superpopulation models, only the regression model which will be used in this thesis.

Let \( \Delta_\varphi \) be the class of all prior distributions \( \delta_\varphi \) satisfying
\[ Y_i = \alpha + \beta X_i + \epsilon_i, \quad i = 1, 2, \ldots, N \quad (2.3.1) \]

with
\[ \mathcal{E}_{\delta_g}(\epsilon_i \mid X_i) = 0 \]
\[ \mathcal{V}_{\delta_g}(\epsilon_i \mid X_i) = \sigma^2 X_i^g \]
\[ C_{\delta_g}(\epsilon_i, \epsilon_j \mid X_i, X_j) = 0 \text{ for } i \neq j. \]

The above model can alternatively be written as follows:
\[ \mathcal{E}_{\delta_g}(Y_i \mid X_i) = \alpha + \beta X_i \]
\[ \mathcal{V}_{\delta_g}(Y_i \mid X_i) = \sigma^2 X_i^g \quad (2.3.2) \]
\[ C_{\delta_g}(Y_i, Y_j \mid X_i, X_j) = 0 \text{ for } i \neq j \]

where \( \mathcal{E}_{\delta_g}, \mathcal{V}_{\delta_g} \) and \( C_{\delta_g} \) denote the conditional expectation, variance and covariance under the prior distributions \( \delta_g \) and \( \alpha, \beta, \sigma^2 \) and \( g \) are unknown constants with \( \sigma^2 > 0 \) and \( 0 \leq g \leq 2 \). In practice \( g \) is found to lie between 1 and 2 and more often close to 2. This is born by empirical studies by Mahalanobis (1944) [34], Smith (1938) [66] and Jesseni (1942) [29].
2.4 The Probability Proportional to Size (PPS) Sampling

One of the methods of utilizing the available auxiliary information at the design stage consists in selecting the units in the sample with probability proportional to a given measure of size (PPS) proposed by Hansen and Hurwitz (1943). The size measure is usually the value of an auxiliary variable \( X \) which is highly positively correlated with the study variable \( Y \). When the sizes of the units are considerably different, the PPS sampling (PPSS) is more appropriate than the SRS which does not take into account the varying size of the units. On the other hand, when the study variable is prima facie correlated with size, it is advantageous to use the PPSS as it assigns probability of selection of units in proportion to their sizes.

Consider again a finite population of \( N \) units,

\[ U \equiv \{U_1, U_2, \ldots, U_N\}. \]

Suppose \( Y \) is the study variate and the size measure is the values of the auxiliary variable \( Z \). We are interested in estimating population total

\[ Y = \sum_{i=1}^{N} Y_i \]  \hspace{1cm} (2.4.1)

of the study variate. The initial selection probabilities are

\[ \{P_i\}, \ i = 1, 2, \ldots, N, \ \text{where} \ P_i = \frac{Z_i}{Z}, \]  \hspace{1cm} (2.4.2)
with

\[ Z = \sum_{i=1}^{N} Z_i \]  

(2.4.3)

is the sum of the sizes for all the population units.

Suppose a sample of size \( n \), \( \{u_1, u_2, \ldots, u_n\} \), where \( u_i \) is the population unit among \( U_1, U_2, \ldots, U_N \) selected at the \( i^{th} \) draw for \( i = 1, 2, \ldots, n \), is selected by the Probability Proportional to Size Sampling With Replacement (PPSWR) Scheme. Then an unbiased estimator of the population total \( Y \) due to Hansen and Hurwitz (1943) is given by

\[ \hat{Y}_{HH} = \sum_{i=1}^{n} \frac{y_i}{p_i} \]  

(2.4.4)

where \( y_i \) is the value of the unit \( u_i \) selected at the \( i^{th} \) draw and \( p_i \) is the corresponding initial selection probability of the unit \( u_i \).

The sampling variance of the strategy (PPSWR: \( \hat{Y}_{HH} \)) is given by

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

(2.4.5)

The expression of \( V(PPSWR : \hat{Y}_{HH}) \), will reduce to zero if \( Y_i \) is exactly proportional to \( P_i \). This result suggests that when \( P_i \) is proportional to some measure of size of the \( i^{th} \) unit \( U_i \) the strategy (PPSWR: \( \hat{Y}_{HH} \)) would be an efficient strategy for the population total \( Y \).

Thus it is expected that the PPSWR sampling would be more efficient than
SRSWR if the size measure \( Z \) is approximately proportional to the study variable \( Y \), that is whenever \( Y \) and \( Z \) are linearly related and the line of regression passes through the origin.

**Comparison between SRS and PPS Sampling**

It may be mentioned that the success of PPS sampling depends heavily on the goodness of measures of size. If the size measure is not even near proportional to the study variable, then the PPS may even be worse than the SRS. On comparison of the variances in the two strategies, viz.,

\[
V(PPSWR : \hat{Y}_{HH}) = \frac{Z}{n} \sum \frac{Y_i^2/Z_i - Y^2}{n} \tag{2.4.6}
\]

and

\[
V(SRSWR : \hat{Y}_{UB}) = \frac{N}{n} \sum \frac{Y_i^2 - Y^2}{n} \tag{2.4.7}
\]

we find that PPSWR would be better than SRSWR if

\[
\sum_i (Z_i - \bar{Z}) \frac{Y_i^2}{Z_i} > 0 \tag{2.4.8}
\]

i.e., if \( Z \) and \( Y^2/Z \) are positively correlated (Raj, 1954) [41]. However this criterion is difficult to apply in practice.

Further the linearity of regression of the study variable \( Y \) on the size measure \( Z \) and the high correlation between them are not sufficient conditions.
for PPSWR to be better than SRSWR. In fact the correlation coefficient between $y$ and $Z$ may be unity and yet the PPSWR sampling may be worse than SRSWR. This was proved by Raj (1954) [41] in the following theorem:

**Theorem 2.4.1.** *If for a finite population $y = a + bz$, so that there is perfect correlation between $y$ and $Z$, the PPS sampling will be less precise than SRSWR if*

$$
\frac{\bar{Z} - \bar{Z}_t}{\bar{Z} \sigma^2_Z} > \frac{b^2}{a^2}
$$

*(2.4.9)*

*where* $\bar{Z} = \frac{1}{N} \sum \left( \frac{1}{Z_t} \right)$ *is the Harmonic mean of $Z$.*

*If $a$ is large, the above inequality is easily satisfied. Thus the regression of the study variable on the size measure should not only be linear but must also pass through close to the origin.*

### 2.4.1 PPSWR Method of Selection in Stratified Sampling

In order to ensure the proper representation of all segments of the population into the sample, the technique of stratification is adopted. When the population under study is heterogeneous, a sample selected by simple random sampling does not guarantee the proper spread of the sample over the whole population. Thus stratified SRSWR is expected to provide more precise estimates than SRSWR for unstratified population. We have seen that the PPSWR sampling is more efficient than SRSWR if the regression of
the study variable on the size measure is linear and passes through a point not very far away from origin. Therefore, the technique of stratification and PPSWR method of selection can both be fruitfully exploited to obtain more precise estimates. If samples are drawn by PPSWR method instead of SRSWR within each stratum, it is called stratified PPSWR (St.PPSWR) sampling.

Suppose a finite population of \( N \) units

\[
U = \{U_1, U_2, \ldots, U_N\}
\]

is divided into \( k \) strata of \( N_i \) units, \( i = 1, 2, \ldots, k \). i.e.,

\[
U = \{U_{11}, \ldots, 1N_1, U_{21}, \ldots, U_{2N_2}, \ldots, U_{k1}, \ldots, U_{kN_k}\}.
\]

For estimating the population total \( Y = \sum_i \sum_j Y_{ij} \) of the study variate \( Y \), a sample of size \( n_i \) is to be selected from \( N_i \) units of the \( i^{th} \) stratum by PPSWR sampling based on a size measure \( Z \) with initial probability

\[
P_{ij} = \frac{Z_{ij}}{Z_i}, \quad j = 1, 2, \ldots, N_i \quad (2.4.10)
\]

with \( Z_i = \sum_j Z_{ij} \) is the total size of the \( i^{th} \) stratum.
The strategy using Hansen-Hurwitz estimator in the stratified PP-SWR sampling design is given by

\[ \hat{Y}_{\text{st,HH}} = \sum_i \sum_j \frac{y_{ij}}{p_{ij}} \]  

(2.4.11)

with sampling variance

\[ V(St.PPSWR : \hat{Y}_{\text{st,HH}}) = \sum_i \frac{1}{n_i} \left\{ \sum_{j=1}^{N_i} \frac{Y^2_{ij}}{P_{ij}} - Y_i^2 \right\} \]  

(2.4.12)

In this thesis, the problem of allocation of sample size to different strata will be studied for HH estimator in the stratified PPSWR sampling design under the superpopulation model.

2.5 Ratio Method of Estimation

Let \( Y_i, X_i \) denote the values of the study variable \( Y \) and the auxiliary variable \( X \) respectively, on \( i^{th} \) population unit for \( i = 1, 2, \ldots, N \). Suppose we want to estimate the ratio \( R = Y/X \), where \( Y \) and \( X \) are the population totals for the variables \( Y \) and \( X \) respectively, on the basis of a sample selected through any given sampling scheme. Let \( \hat{Y} \) and \( \hat{X} \) be unbiased estimators of \( Y \) and \( X \) respectively. Then an estimator of the ratio \( R \) is given by

\[ \hat{R} = \frac{\hat{Y}}{\hat{X}}. \]
Similarly, a ratio estimator of the population total $Y$ is given by

$$\hat{Y}_R = \hat{R}X = \frac{\hat{Y}}{X} \hat{X} X \tag{2.5.1}$$

provided information on the total $X$ of the auxiliary variable is available from some other source.

The ratio estimator is biased. For the derivations of the bias and MSE of the ratio estimator (2.5.1) let us write

$$\hat{Y} = Y(1 + e) \text{ and } \hat{X} = X(1 + e')$$

where

$$E(e) = E(e') = 0 .$$

Then

$$E(\hat{Y}) = Y E[(1 + e)(1 + e')^{-1}]$$

Assuming

$$(i) \quad |e'| < 1 \quad \text{i.e.,} \quad \left| \frac{\hat{X} - X}{X} \right| < 1.$$
and neglecting the terms with powers of $e$, $e'$ greater than 2 in the expansion of infinite series, i.e., upto the second degree of approximation, it would be easy to derive the formula for Mean Square Error (MSE) and Bias [Murthy (1967) [39]], which are given below:

The approximate bias of $\hat{Y}_R$ is given by

$$B(\hat{Y}_R) = \frac{1}{X} [R \, V(\hat{X}) - Cov(\hat{X}, \hat{Y})]. \tag{2.5.2}$$

The bias $B(\hat{Y}_R)$ up to the second order of approximation decreases with increasing sample size for most of the sampling designs commonly used in practice. The approximate expression of bias vanishes for $R = \frac{Cov(\hat{X}, \hat{Y})}{V(\hat{X})}$, which is satisfied when the regression line of $\hat{Y}$ on $\hat{X}$ is a straight line passing through the origin. In such a case, the MSE of $\hat{Y}_R$ which is the same as the variance $V(\hat{Y}_R)$, is approximately given by

$$V(\hat{Y}_R) = V(\hat{Y}) - 2 \, R \, Cov(\hat{X}, \hat{Y}) + R^2 \, V(\hat{X}). \tag{2.5.3}$$

Though, the ratio estimator is biased yet the MSE and the square of bias of the ratio estimator taken together is less than the variance of the usual unbiased estimator. The above expressions of the bias and MSE of $\hat{Y}_R$ are derived under the following two assumptions:

(i) $\left| \frac{\hat{X} - X}{\hat{X}} \right| < 1$, i.e., $\hat{X}$ lies between 0 and $2X$, which is likely to be valid only for large samples.
(ii) terms of degree greater than 2 in \((e, e')\) in the expansion for \((1 + e) (1 + e')^{-1}\) can be neglected.

These assumptions are likely to be valid only for large samples. It was shown by Koop (1972) that the first condition will always be satisfied for \(n > N/2\) which is a severe restriction on the sample size in practice. However, whether this condition is satisfied or not, which depends on the nature of the survey population, the above expressions of Bias and MSE are useful approximations in practice.

When simple random sampling with replacement (SRSWR) is used, the variance of the ratio estimator, \(V(\hat{Y}_{SR})\) in (2.5.3) reduces to [see Murthy(1967) [39]]

\[
V(\hat{Y}_{SRWR}) = \frac{\sigma_u^2}{n} = \frac{1}{n} \frac{1}{N} \sum_{i=1}^{N} U_i^2
\]  

(2.5.4)

where

\[
\sigma_u^2 = \sigma_y^2 - 2R \sigma_{yx} + R^2 \sigma_x^2
\]

is the variance of the transformed variable

\[U_i = Y_i - RX_i.\]

Thus
Similarly for simple random sampling without replacement (SRSWOR) the variance of the ratio estimator becomes [see e.g., Cochran (1977) [8] and Sukhatme et al. (1984) [76]],

\[
V(\hat{Y}_{R,W_R}) = \frac{1}{n} \frac{1}{N} \left( \sum_{i=1}^{N} Y_i^2 - 2R \sum_i Y_i X_i + R^2 \sum_i X_i^2 \right). \tag{2.5.5}
\]

Similarly for simple random sampling without replacement (SRSWOR) the variance of the ratio estimator becomes [see e.g., Cochran (1977) [8] and Sukhatme et al. (1984) [76]],

\[
V(\hat{Y}_{R,WOR}) = \left( \frac{1-f}{n} \right) S_u^2
\]

\[
= \left( \frac{1-f}{n(N-1)} \right) \left( \sum_i Y_i^2 - 2R \sum_i Y_i X_i + R^2 \sum_i X_i^2 \right). \tag{2.5.6}
\]

where \( f = \frac{n}{N} \) is the sampling fraction.

### 2.5.1 Ratio Estimation in Stratified Sampling

The technique of stratification produces gains when the population is heterogenous compared to unstratified sampling. On the other hand ratio estimator is more precise than the usual unbiased estimator when the regression line of the study variate on the auxiliary variable is linear and passes through the origin. Therefore it would be worth to combine the stratification of the population and use of the ratio estimation for enhancing the precision of the estimates obtained through sample survey.

In stratified sampling, a ratio estimator can be used in two ways. First, is to make ratio estimate for every stratum total and then add these
stratum estimates to build-up the overall estimate of the population total. This is called Separate Ratio (SR) Estimator (SRE). The second type of Ratio Estimator in stratified sampling is due to Hansen, Hurwitz and Gurney (1946) which uses estimate of the overall ratio to construct the estimator for the population total. This is called Combined Ratio (CR) Estimator (CRE).

**Separate Ratio Estimator**

It is assumed that the regression of the study variate $Y$ on the auxiliary variate $X$ is linear passing through the origin and the stratum total of the auxiliary variable for every stratum is available. Further, if the sample size in each stratum is large enough so that the assumption $|\frac{\bar{X}_i}{\bar{X}} - 1| < 1$ holds and hence $(1 - \frac{\bar{X}_i}{\bar{X}})^{-1}$ can be expanded.

The ratio estimator of the $i^{th}$ stratum total from (2.5.1) is given by

$$\hat{Y}_{i,R} = \hat{R}_i X_i .$$

Therefore the separate ratio estimator is given by

$$\hat{Y}_{SR} = \sum_{i=1}^{k} \hat{Y}_{i,R} = \sum_{i=1}^{k} \hat{R}_i X_i . \quad (2.5.7)$$

If the sample size is large in each stratum, the bias of $\hat{Y}_{SR}$ given by (2.5.2) is given by
\[ B(\hat{Y}_{SR}) = \sum_{i=1}^{k} \frac{1}{X_i} \left[ R_i \, V(\hat{X}_i) - Cov (\hat{X}_i, \hat{Y}) \right]. \]

The approximate variance of \( \hat{Y}_{SR} \) up to the second order of approximation using (2.5.3) is given by [see Murthy (1967) [39]]

\[ V(\hat{Y}_{SR}) = \sum_{i=1}^{k} V(\hat{Y}_{i,n}) \]
\[ = \sum_{i=1}^{k} \left[ V(\hat{Y}_i) + \hat{R}_i V(\hat{X}_i) - 2\hat{R}_i \, Cov (\hat{Y}_i, \hat{X}_i) \right]. \] (2.5.8)

When SRSWR is adopted within each stratum, the variance \( V(\hat{Y}_{SR}) \) in (2.5.8) using (2.5.4) reduces to

\[ V(\hat{Y}_{SR}) = \sum_{i=1}^{k} \frac{N_i^2 \, \sigma_i^2(u)}{n_i} = \sum_{i=1}^{k} \frac{N_i^2}{n_i} \frac{1}{N_i} \sum_{j} U_{ij}^2 \] (2.5.9)

where

\[ \sigma_i^2(u) = \frac{1}{N_i} \sum_{j} (Y_{ij} - R_i X_{ij})^2 \] (2.5.10)

\[ = \sigma_i^2(y) - 2R_i \, \sigma_i(x,y) + R_i^2 \, \sigma_i^2(x) \]

is the variance of the transformed variable

\[ U_{ij} = Y_{ij} - R_i X_{ij}. \]

Therefore,

\[ V(\hat{Y}_{SR,WL}) = \sum_{i=1}^{k} \frac{N_i^2}{n_i} \frac{1}{N_i} \left( \sum_{j} Y_{ij}^2 - 2R_i \sum_{j} X_{ij} Y_{ij} + R_i^2 \sum_{j} X_{ij}^2 \right). \] (2.5.11)

32
On the other hand if SRSWOR is adopted within each stratum the variance from (2.5.8) using (2.5.6) is given by [see Cochran (1977) [8] and Sukhatme et al. (1984)], [76]

\[
V(\hat{Y}_{SR\text{-}WR}) = \sum_{i=1}^{k} N_{i}^{2} \left( \frac{1 - f_{i}}{n_{i}} \right) S_{i}^{2}(u)
\]

\[
= \sum_{i=1}^{k} N_{i}^{2} \left( \frac{1 - f_{i}}{n_{i}} \right) \frac{1}{N_{i} - 1} \left( \sum_{j} Y_{ij}^{2} - 2R_{i} \sum_{j} X_{ij}Y_{ij} + R_{i}^{2} \sum_{j} X_{ij}^{2} \right)
\]

(2.5.12)

where \( f_{i} = n_{i}/N_{i} \) is the sampling fraction in the \( i^{th} \) stratum.

The Combined Ratio Estimator

Suppose \( \hat{Y}_{st} \) and \( \hat{X}_{st} \) are the unbiased estimators of the population totals of \( Y \) and \( X \) respectively from stratified sample than the combined ratio estimator [Hansen, Hurwitz and Gurney (1946)] of the population total of \( Y \) is given by

\[
\hat{Y}_{CR} = \frac{\hat{Y}_{st}}{\hat{X}_{st}} X.
\]

Since
\[ V(\hat{Y}_{st}) = \sum_i V(\hat{Y}_i) \]

\[ V(\hat{X}_{st}) = \sum_i V(\hat{X}_i) \]

\[ \text{Cov}(\hat{X}_{st}, \hat{X}_{st}) = \sum_i \text{Cov}(\hat{Y}_i, \hat{X}_i). \]

For large samples, the bias of this estimator is given by

\[ B(\hat{Y}_{CR}) = \left( RV(\hat{X}_{st}) - \text{Cov}(\hat{X}_{st}, \hat{Y}_{st}) \right) \] (2.5.13)

vanishes and its approximate variance is given by

\[ V(\hat{Y}_{CR}) = V(\hat{Y}_{st}) + R^2 V(\hat{X}_{st}) - 2R \text{Cov}(\hat{X}_{st}, \hat{Y}_{st}) \]

\[ = \sum_i \left[ V(\hat{Y}_i) + R^2 V(\hat{X}_i) - 2R \text{Cov}(\hat{X}_i, \hat{Y}_i) \right]. \] (2.5.14)

Note that this estimate does not require a knowledge of the individual stratum total \(X_i\)'s but only overall population total \(X\).

When SRSWR is adopted in each stratum, the variance \(V(\hat{Y}_{CR})\) from (2.5.14) is given by

\[ V(\hat{Y}_{CR,WR}) = \sum_i \frac{N_i^2 \sigma_i^2(u)}{n_i} \]

\[ = \sum_i \frac{N_i^2}{n_i} \left\{ \sigma_i^2(y) - 2R \sigma_i(x, y) + R^2 \sigma_i^2(u) \right\} \] (2.5.15)
where

\[ \sigma_i^2(v) = \frac{1}{N_i} \sum_j (V_{ij} - \bar{V}_i)^2 \]

is the variance in the \( i^{th} \) stratum of the transformed variable

\[ V_{ij} = Y_{ij} - R X_{ij} . \]

On the other hand for SRSWOR within each stratum, the variance (2.5.14) reduces to

\[
V(Y_{cr, wor}) = S_f(v) = \sum N_i \left( \frac{1}{n_i} \right) \left\{ S_i^2(y) - 2R \, S_i(x, y) + R^2 \, S_i^2(x) \right\}
\]

where

\[ S_i^2(v) = \frac{N_i}{N_i - 1} \sigma_i^2(v) \]
Comparison of the Combined and Separate Ratio Estimators

The bias of CRE \( \hat{Y}_{CR} \) is expected to be much less than that of SRE \( \hat{Y}_{SR} \) because the ratios \( \{ R_i \} \) may not be much different from \( R \) but \( \{ X_i \} \) would be much smaller than \( X \). Thus CRE is much less subject to the risk of bias than SRE.

Now

\[
V(\hat{Y}_{CR}) - V(\hat{Y}_{SR}) = \sum_i (R_i^2 - R_i^2) \hat{V}(\hat{X}_i) - 2 \sum_i (R - R_i) \text{Cov}(\hat{X}_i, \hat{Y}_i).
\]

(2.5.17)

The last term on the right is usually small and it would vanish if the regression if \( Y \) on \( X \) is a straight line through the origin in each stratum. This shows that if the stratum ratios do not differ among themselves, the combined ratio estimator \( \hat{Y}_{CR} \) should be preferred to the separate ratio estimator. However, if \( R_i \)’s differ much among themselves, the separate ratio estimator in each stratum is expected to be more precise if the sample in each stratum is large enough so that the approximate formula for \( V(\hat{Y}_{SR}) \) is valid.

Further, we note that for obtaining the expressions for the bias and the variance of \( \hat{Y}_{CR} \); it is sufficient to assume that \( \frac{\hat{X}_i - X_i}{\hat{X}_i} < 1 \), whereas for \( \hat{Y}_{SR} \); this condition is required for each stratum i.e., \( \frac{\hat{X}_i - X_i}{X_i} < 1 \). Thus, the expressions of the bias and variance for \( \hat{Y}_{CR} \) are valid if the total sample size
is large enough, the corresponding expressions derived for $\hat{Y}_{SR}$ will be valid only when the sample size in each stratum is sufficiently large. Thus, with only small sample in each stratum, the combined ratio estimator should be preferred to separate ratio estimator unless there is good empirical evidence to the contrary.

We now state two theorems without proof on comparison of separate and combined ratio estimators.

**Theorem 2.5.1.** Let

\[
\hat{Y}_{CR} = \sum_{i=1}^{k} \frac{\hat{Y}_i}{\hat{X}_i} X
\]  

(2.5.18)

and

\[
\hat{Y}_{SR} = \sum_{i} \frac{\hat{Y}_i}{\hat{X}_i} X
\]  

(2.5.19)

be the combined and separate ratio estimator of $Y$ respectively based on stratified sampling with any sampling design in each stratum. Then, to the second degree approximation

\[
B(\hat{Y}_{CR}) = 0 \neq \left| B\left(\hat{Y}_{SR}\right) \right|
\]

\[
M(\hat{Y}_{CR}) = M\left(\hat{Y}_{SR}\right)
\]

provided $R_i$'s are equal

and

37
\[ B(\hat{Y}_{SR}) = 0 \neq |B(\hat{Y}_{CR})| \]

\[ M(\hat{Y}_{SR}) = M(\hat{Y}_{CR}) \]

provided \( \kappa_i \) for all \( i \),

where

\[ \kappa_i = \frac{\text{Cov}(\hat{Y}_i, \hat{X}_i)}{V(\hat{X}_i)} \cdot \frac{X_i}{Y_i} \]

From the above theorem it is clear that \( \hat{Y}_{CR} \) will be preferable to \( \hat{Y}_{SR} \) whenever the stratum ratios \( R_i \)'s do not differ significantly from one another, whereas \( \hat{Y}_{SR} \) is preferable to \( \hat{Y}_{CR} \) whenever the individual regression lines in each stratum pass through a point close to the origin.

**Theorem 2.5.2.** [Reddy (1975)] Let \( \hat{Y}_i \) and \( \hat{X}_i \) be unbiased estimators of population totals \( Y_i \) and \( X_i \) of the \( i^{th} \) stratum respectively based on a simple random sample without replacement (SRSWOR) of size \( n_i \) drawn from the \( i^{th} \) stratum, \( i = 1, 2, \ldots, k \).

Let

\[ \hat{Y}_{CR} = \frac{\sum_{i=1}^{k} \frac{\hat{Y}_i}{\hat{X}_i} X}{\sum_{i} \hat{X}_i} \]

(2.5.20)
and

\[ \hat{Y}_{SR} = \sum_{i=1}^{k} \hat{Y}_i X_i \]  \hspace{1cm} (2.5.21)

where \( \sum_i X_i = X \). Then under the superpopulation model \( \Delta_g \), there exists a \( g_0 \) in \((0,1)\) such that

\[ \varepsilon_{\delta_g} M(\hat{Y}_{SR}) < \varepsilon_{\delta_g} M(\hat{Y}_{CR}) \text{ for } g \geq g_0. \]  \hspace{1cm} (2.5.22)

The choice between SRE and CRE depends mainly on the stratum ratios whether they differ among themselves or not. Either past survey data, if available or a pilot survey may be useful for collecting the information on stratum ratios and their equality may be tested. For statistical tests on equality of slopes we refer Rao, [P.S.R.S. Rao (2000)] [47].

The problem of allocation of sample size to strata for both the SRE and CRE in stratified SRS sampling will be considered in chapters 6 and 7 respectively.