Two Phase Sampling

Two Phase Sampling

By

Zahoor Ahmad, Muhammad Qaiser Shahbaz
and Muhammad Hanif

CAMBRIDGE
SCHOLARS

PUBLISHING

Two Phase Sampling,
by Zahoor Ahmad, Muhammad Qaiser Shahbaz and Muhammad Hanif

This book first published 2013

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-4595-7, ISBN (13): 978-1-4438-4595-3

CONTENTS

Section-I: Single Phase Sampling with Auxiliary Variable

Section-II: Two Phase Sampling with Auxiliary Variable

Section-III: Single and Two Phase Sampling with Multi-auxiliary Variable

Section-IV: Single and Two Phase Sampling with Auxiliary Attributes

PREFACE

This book attempts to give an up to date account and the developments of two-phase sampling using simple random sampling without replacement at each phase. This book also provides a comprehensive review of the previous developments of various researchers in single and two-phase sampling, along with recent developments of the present authors. This book may be used as reference and as a textbook for undergraduate and post-graduate students. This book consists of 9 chapters divided into four sections.

Section one deals with single phase sampling for single and two auxiliary variables, while section two deals with two-phase sampling using single and two auxiliary variables. In section three we have given estimators which are based on multi-auxiliary variables, and section four deals with estimators based on attributes.

We are thankful to Dr. Ken Brewer, Dr. Munir Ahmad, and Dr. A. Z Memon for the clarification of some concepts. We are indebted to Dr. Inam-ul-Haq, Dr Hammad Naqvi, Ms Munaza Bajwa, Ms Asma Tahir and Ms Saria Bano for checking the derivations of mean square errors, some estimators, and for conduct of empirical studies. In general we are thankful to all the M.Phil. and Ph.D. students of National College of Business and Economics, Lahore, for their help at various stages.

—Zahoor Ahmad, M. Qaiser Shahbaz and Muhammad Hanif

CHAPTER ONE

BASIC SAMPLING THEORY

1.1 Introduction

Most survey work involves sampling from finite populations. There are two parts of any sampling strategy. First there is the selection procedure, the manner in which sampling units are to be selected from a finite population. Second there is an estimation procedure, which prescribes how inferences are to be made from sample to population. These inferences may be either enumerative or analytical. Enumerative inference seeks only to describe the finite population under study, whereas analytical inference attempts to explain the underlying distributional and functional characteristics of a population. Enumerative inference typically concerns the estimation of some parameters of a population such as means, totals, proportion and ratios. Viewing the same population analytically we might be interested to regress household income on such variables like number of employed adults, educational level of household head etc. The fitted regression model is an explanatory model. Analytical inference consists of appropriate specification of a model which adequately describes the sample, and hence the population from which it was selected. It is customary to distinguish between enumerative and analytical inference in terms of complexity of the population characteristics of estimator. Estimating a mean, total, etc., is regarded as an enumerative problem, whereas estimating a regression or correlation coefficient is seen as an analytical one, but if the mean in question is a parameter of a simple explanatory model, the inference is enumerative.

Analytical and enumerative inferences used by certain models provide their own probability structure, which guides directions for further inferences. The same methodology is used in other areas of statistics. For enumerative inference, a quite different probability structure is used; it depends on the manner in which the sample is selected. This is the classical finite population sampling inference developed by Neyman (1934), who based his results on Gram (1883) and Bowley (1913) [Brewer and Hanif (1983)].

The sampling theory attracted the most attention as compared with other fields of statistics in the post-war period. This was probably due to methods used in sampling which have many practical applications. It is sufficient to recall that the use of sampling theory has entirely changed the work of data collection and seems to be the reason why most of the names among contemporary statisticians have devoted much of their time to some problems in the theory and practices of sample surveys.

Systematic interest in the use of sampling theory appeared towards the end of the last century when Kiaer (1890) used the representative method for collecting data independently of the census. In 1901 he also demonstrated empirically that stratification could provide good estimates of finite population totals and means. On the recommendation of Kiaer, the International Statistical Institute in 1903 adopted stratified sampling with proportional allocation as an acceptable method of data collection. In line with the present thinking, the selection procedure used by Kiaer and others deals with a method of drawing a representative sample from the population. The primary difference in the earlier use of the sampling methods lay in the selection of the sample. Earlier it was thought that a purposive selection based on the sampler's knowledge of the population with regard to closely correlated characteristic was the best way of getting a sample that could be considered representative of the population. Neyman (1934) published his revolutionary paper in which he made it clear that random selection had its basis on a sound scientific theory, which definitely gave sample survey the character of an objective research tool and made it possible to predict the validity of the survey estimates. Actually in this sense this is marked as the beginning of a new era. Neyman (1934) suggested optimum allocation in stratified sampling. Nowadays, sampling theory has been so developed in its application that it is widely used in all the fields of life and the use of this theory is proved to be fruitful in the underdeveloped areas as well.

1.2 The Use of Auxiliary Information

The history of use of auxiliary information in survey sampling is as old as the history of survey sampling. Graunt (1662) seems to be the first statistical scientist who estimated the population of England based on classical ratio estimate using auxiliary information. At some time in the mid-1780s (the exact date is difficult to establish), the eminent mathematician Laplace (1783) started to press the ailing French government to conduct an enumeration of the population in about 700 communes scattered over the Kingdom (Bru, 1988), with a view to estimating the total population of

France. He intended to use for this purpose the fact that there was already a substantially complete registration of births in all communes, of which there would then have been of the order of 10,000. He reasoned that if he also knew the populations of those sample communes, he could estimate the ratio of population to annual births, and apply that ratio to the known number of births in a given year, to arrive at what we would now describe as a ratio estimate of the total French population (Laplace, 1783, 1814a and 1814b). For various reasons, however, notably the ever-expanding borders of the French empire during Napoleon's early years, events militated against him obtaining a suitable total of births for the entire French population, so his estimated ratio was never used for its original purpose (Bru, 1988; Cochran, 1977, Hald, 1998; Laplace, 1814a and 1814b, p. 762). He did, however, devised an ingenious way for estimating the precision with which that ratio was measured. This was less straightforward than the manner in which it would be estimated today but, at the time, it was a very considerable contribution to the theory of survey sampling.

The works of Bowley (1926) and Neyman (1934, 1938) are the foundation stones of modern sampling theory. These two authors presented use of stratified random sampling and they also put forward theoretical criticism on non-random sampling (purposive sampling). These two classical works may perhaps be referred to as the initial work in the history of survey sampling where use of auxiliary information has been demonstrated. It was a general intuitive of the survey statisticians during the 1930s that the customary method of estimating the population mean or total of a variable of interest (estimand), say y, may be improved to give higher precision of estimation if the information supplied by a related variable (auxiliary variable, supplementary variable, concomitant variable, ancillary variable or benchmark variable), say x, is incorporated intelligibly in the estimation procedure. The work of Watson (1937) and Cochran (1940, 1942) was also the initial work making use of auxiliary information in devising estimation procedures leading to improvement in precision of estimation. Hansen and Hurwitz (1943) were the first to suggest the use of auxiliary information in selecting the population with varying probabilities. In most of the survey situations the auxiliary information is always available. It may either be readily available or may be made available without much difficulty by diverting a part of the survey resources. The customary sources of obtaining relevant auxiliary information on one or more variables are various: i.e. census data, previous surveys or pilot surveys. The auxiliary information may in fact be available in various forms. Some of them may be enumerated as follows.

For details see Tripathi (1970) and Das (1988). In brief, four ways are given as:

i) The values of one or more auxiliary variables may be available a priori only for some units of a finite population,
ii) Values of one or more parameters of auxiliary variables, i.e. population mean(s), population proportion(s), variance, coefficient(s) of variation, Coefficient of skewness may be known. In other words one or more parameters are known,
iii) The exact values of the parameters are not known but their estimated values are known,

And

iv) The values of one or more auxiliary variables may be known for all units of a finite population.

It is worth mentioning here that in addition to the information on auxiliary variable *x* (in one form or another), information about the estimated variable *y* may also be available in summary form in some cases. For example coefficient of variation C_y of *y* may be known exactly [Searle (1964)], or approximate mean square error may be available.

Various research works during the last 70 years confirmed the view that whatever the form of availability of auxiliary information, one may always utilize it to devise sampling strategies which are better (if not uniformly then at least in a part of parametric space) than those in which no auxiliary information is utilized. The method of utilization of auxiliary information depends on the form in which it is available.

As mentioned by Tripathi (1970, 1973, 1976), the auxiliary information in survey sampling may be utilized in the following four ways:

i) The auxiliary information may be used at the planning or designing stage of a survey, i.e. in stratifying the population, the strata can be made according to auxiliary information.
ii) The auxiliary information may be used for the purpose of estimation, i.e. ratio, regression, difference and product estimators etc.
iii) The auxiliary information may be used while selecting the sample, i.e. probabilities proportional to size sampling.

iv) The auxiliary information may be used in mixed ways, i.e. combining at least two of the above.

The classical ratio and regression estimators [Cochran (1940, 42)], difference estimator [Hansen et al. (1953)] and product estimator [Robson (1957), Murthy (1964)] for population mean of estimand variable, which are based on the knowledge of population mean $\overline{X}$ of an auxiliary variable *X*, are well known in the sample survey literature and their detailed study in case of simple random sampling without replacement, and stratified random sampling are available in various books, i.e. [Yates (1960), Deming (1960), Kish (1965), Murthy (1967), Raj (1968), Cochran (1977), and Sukhatme et al. (1984), etc.]. Later these classical estimators were extended to various sampling designs by several researchers. The ratio and regression estimators being biased, the efforts were made by various researchers to obtain unbiased ratio and regression type estimators during the 1950s and 1960s. Hartly and Ross (1954) defined an unbiased ratio estimator based on simple random sampling using the knowledge of auxiliary variable. Basically four methods have been used to obtain unbiased or almost unbiased estimators.

i) Bias correction using the knowledge of $\overline{X}$ [Hartley and Ross (1954), Beale (1962) and Tin (1965)]
ii) Jack-knife technique [Quenouille (1956), Durbin (1959) and Mickey (1959).
iii) Mahalanobis's technique of interpenetrating sub-samples. [Murthy and Nanjamma (1959)]
iv) Varying probabilities technique based on the information of X. [Lahiri (1951), Sen (1952), Midzuno (1952) and Ikeda (1952) reported by Midzuno (1952)].

Similar methods were used to obtain an unbiased regression estimator [Williams (1963), Robson (1957)]. Srivenkataramana (1980) and Srivenkataramana and Tracy (1979, 80, 81) were the first to introduce transformation of auxiliary variable X to change negative correlation situation into a positive one and vice-versa, giving duals to ratio estimator. Tripathi and Singh (1988) gave a generalized transformation capable of dealing with positive and negative correlation situations simultaneously and derived a wide class of unbiased product type estimators which also includes the unbiased product type estimators due to T.J. Rao (1983,87), Singh et al. (1987) and several others.

The study of ratio and regression estimators based on super-population models have been made by several authors during the 1960s and 1970s. In this context we refer to Brewer (1963), P. S. R. S. Rao (1969), Royall (1970), Cassel et al. (1977), etc. The performance of ratio estimator based on small samples has been studied by P. S. R. S. Rao and J. N. K. Rao (1971) and others.

The use of multivariate auxiliary information in defining ratio, difference, regression, product, and ratio-cum-product type estimators etc. for population mean $\bar{Y}$, under various situations, has been considered by Olkin (1958), Raj (1965), Srivastava (1965, 1966, 71), Shukla (1965, 1966), Rao and Mudholkar (1967), Singh (1967), Khan and Tripathi (1967), Tripathi (1970, 1976), etc. Tripathi (1987) unified most of the results by considering a class of estimators based on general sampling design and multivariate information.

The problem of estimation of population ratio $R = \frac{Y}{X}$ or $R = \frac{Y}{Z}$ using auxiliary variable did not attract much attention during the 1950s. The work relating to the use of auxiliary information in estimating α was initiated by J. N. K. Rao (1957), Singh (1965, 1967, 1969), Rao and Pereira (1968), and Tripathi (1980). This work was followed by several authors. The use of auxiliary information in the form of coefficient of variation C_x of an auxiliary variable X for estimation of Y was considered first by Das and Tripathi (1980). Das and Tripathi (1978) also initiated the work related to the use of auxiliary information for estimating the finite population variance σ_y^2. Das and Tripathi (1979) considered the use of σ_x^2. Srivastava and Jhajj (1981) and Das and Tripathi (1981) considered the simultaneous utilization of $\bar{X}$ and σ_x^2 for estimation of $\bar{Y}$. Srivatava and Jhajj (1980) also considered the use of $\bar{X}$ and σ_x^2 for estimating σ_y^2. The use of more than one auxiliary variable in defining the selection probabilities to select the population units with varying probabilities and with replacement was considered by Maiti and Tripathi (1976), and Agarwal and Singh (1980).

1.3 Difference between Multi-stage and Multi-phase Sampling

The differences between multistage and multiphase sampling are given below:

- The first difference is that in multistage sampling the natures (and most specifically the sizes) of the population units differ from stage to stage; whereas in multiphase sampling there is only one kind of population unit, and each such unit could end up being selected, either just at the first phase or at one or more consecutive subsequent phases also.
- The second difference, which builds on the first, is that the units at any stage are larger than units from the stage immediately following. (They must be larger, being actually comprised of them! and consequently, they themselves also make up units from the immediately previous stage.)

For example, in an Australian Labour Force Survey, the first stage units were typically Local Government Areas, containing thousands of dwellings; the second stage units were typically Census Collectors' Districts containing typically two or three hundred dwellings each; the third stage units were typically street blocks, each usually containing a few dozen dwellings; and the fourth stage units were typically the individual dwellings themselves (though occasionally they were houses containing more than one household; and therefore, by definition, more than one dwelling). By contrast, in multi-phase sampling, any unit of the population can become a first phase sample unit, and, if selected at the first stage, can also be a second phase sample unit, and so on.

- The third difference is that in multistage sampling, information about the actual survey variables is gathered only from the units selected at the final stage, e.g. from all the people living in a selected dwelling, but not from the remainder of the people in a selected Block (or Collector's District or Local Government Area). By contrast in multiphase sampling, we typically collect only a little information from (or, more often, a little information about) first phase sample units. It has to be relatively cheap information, available perhaps from administrative records; otherwise there would be no point in using a multiphase sample at all. It also has to be information that is correlated with (in practice usually more or less proportional to) the survey variables of interest.
- A fourth commonly occurring difference, though not an essential one, is that multistage sampling can involve three or sometimes even four stages, though it is quite uncommon to have more than two phases. For this reason, another expression for two-phase sampling is "double sampling".

- The last difference we can think of is that one usually has to work one's way up the stages by calculating (say for a three-stage sample) first, third stage sample estimates for the relevant two-stage sample units; next, use those estimates to arrive at first stage sample estimates, and finally use those estimates in turn to estimate figures for the entire population. In multiphase sampling, only the entire population is estimated. It can be viewed as being estimated (rather poorly) on the basis of the lowest phase sample only, rather better by incorporating information from the next phase up, and so on, with the most accurate estimate being made on the basis of information from all the phases of selection together. Each phase of the sample, other than the last, would therefore be supplying only an adjustment factor of the order of unity.

1.4 Notations

Consider a population of N units. Associated with the Ith unit are the variables of main interest Y_I and the auxiliary variables X_I and Z_I for $I = 1, 2, 3, ..., N$. Let $\bar{X}$, $\bar{Z}$ and $\bar{Y}$ be the population mean of variables *X*, *Z* and *Y* respectively. Further let S_x^2, S_z^2 and S_y^2 be corresponding variances. The coefficient of variation of these variables will be denoted by $C_x^2 = \left(S_x^2 / \bar{X}^2\right)$, $C_y^2 = \left(S_y^2 / \bar{Y}^2\right)$ and $C_z^2 = \left(S_z^2 / \bar{Z}^2\right)$ respectively. Also ρ_{xy}, ρ_{xz} and ρ_{yz} will represent population correlation coefficients between *X* and *Y, X* and *Z,* and *Y* and *Z* respectively.

The first phase sample means based on n_1 units are denoted by $\bar{y}_1$, $\bar{x}_1$ and $\bar{z}_1$ etc. and second phase sample means are denoted by $\bar{y}_2, \bar{x}_2$ and $\bar{z}_2$. We will also use $\theta_1 = n_1^{-1} - N^{-1}$ and $\theta_2 = n_2^{-1} - N^{-1}$ so that $\theta_2 > \theta_1$. For notational purpose we will assume that the mean of estimand (estimated variable) and auxiliary variables can be approximated from their population means so that $\bar{x}_h = \bar{X} + \bar{e}_{x_h}$, where $\bar{x}_h$ is sample mean of auxiliary variable *X* at *h*–th phase for *h* = 1 and 2. Similar notation will be used for other quantitative auxiliary variables. For qualitative auxiliary variables we will use $p_h = P + \bar{e}_{\tau_h}$. In using the relation between sample and population mean of auxiliary variable we will assume that $\left|\bar{e}_{x_h}\right|$ is much smaller as compared with $\left|\bar{X}\right|$. For variable of interest we will use $\bar{y}_h = \bar{Y} + \bar{e}_{y_h}$ with usual assumptions.

In this monograph we will use following expectations for deriving the mean square error of estimators which are based upon quantitative auxiliary variables:

$$\left.\begin{aligned}
&E\left(\bar{e}_{x_1}^2\right) = \theta_1 \bar{X}^2 C_x^2 \ ; \ E\left(\bar{e}_{z_1}^2\right) = \theta_1 \bar{Z}^2 C_z^2 ; \ E\left(\bar{e}_{y_1}^2\right) = \theta_1 \bar{Y}^2 C_y^2 \\
&E\left(\bar{e}_{x_2}^2\right) = \theta_2 \bar{X}^2 C_x^2 \ ; \ E\left(\bar{e}_{z_2}^2\right) = \theta_2 \bar{Z}^2 C_z^2 ; \ E\left(\bar{e}_{y_2}^2\right) = \theta_2 \bar{Y}^2 C_y^2 \\
&E\left(\bar{e}_{x_1}\bar{e}_{x_2}\right) = \theta_1 \bar{X}^2 C_x^2 ; \ E\left(\bar{e}_{z_1}\bar{e}_{z_2}\right) = \theta_1 \bar{Z}^2 C_z^2 \\
&E\left(\bar{e}_{x_1}\bar{e}_{y_1}\right) = \theta_1 \bar{X}\bar{Y} C_x C_y \rho_{xy} \ ; \ E\left(\bar{e}_{x_1}\bar{e}_{y_2}\right) = E\left(\bar{e}_{x_2}\bar{e}_{y_1}\right) = \theta_1 \bar{X}\bar{Y} C_x C_y \rho_{xy} \\
&E\left(\bar{e}_{x_2}\bar{e}_{y_2}\right) = \theta_2 \bar{X}\bar{Y} C_x C_y \rho_{xy} ; \ E\left(\bar{e}_{x_1}\bar{e}_{z_1}\right) = \theta_1 \bar{X}\bar{Z} C_x C_z \rho_{xz} \ \ etc \\
&E\left(\bar{e}_{x_1} - \bar{e}_{x_2}\right)^2 = \left(\theta_2 - \theta_1\right) \bar{X}^2 C_x^2 \ \ etc \\
&E\left\{\bar{e}_{y_1}\left(\bar{e}_{x_1} - \bar{e}_{x_2}\right)\right\} = E\left\{\bar{e}_{z_1}\left(\bar{e}_{x_1} - \bar{e}_{x_2}\right)\right\} = 0 \\
&E\left\{\bar{e}_{y_2}\left(\bar{e}_{x_1} - \bar{e}_{x_2}\right)\right\} = \left(\theta_1 - \theta_2\right) \bar{X}\bar{Y} C_x C_y \rho_{xy} \ \ etc
\end{aligned}\right\} \tag{1.4.1}$$

In case of several auxiliary variables; say *q*; the sample mean of *i*th auxiliary at *h*th phase will be denoted by $\bar{x}_{(i)h} = \bar{X}_i + \bar{e}_{x_{(i)h}}$. The expectations in (1.4.1) are modified as:

$$\left.\begin{aligned}
&E\left(\bar{e}^2_{x_{(i)1}}\right)=\theta_1\bar{X}_i^2C^2_{x_i}\ ;\ E\left(\bar{e}^2_{x_{(i)2}}\right)=\theta_2\bar{X}_i^2C^2_{x_i}\\
&E\left(\bar{e}_{x_{(i)1}}\bar{e}_{x_{(i)2}}\right)=\theta_1\bar{X}_i^2C^2_{x_i}\ ;\ E\left(\bar{e}_{x_{(i)1}}\bar{e}_{x_{(k)2}}\right)=\theta_1\bar{X}_i\bar{X}_kC_{x_i}C_{x_k}\rho_{ik}\ ;\\
&E\left(\bar{e}_{x_{(i)1}}\bar{e}_{y_1}\right)=\theta_1\bar{X}_i\bar{Y}C_{x_i}C_y\rho_{yx_i}\ ;\ E\left(\bar{e}_{x_{(i)2}}\bar{e}_{y_2}\right)=\theta_2\bar{X}_i\bar{Y}C_{x_i}C_y\rho_{yx_i}\\
&E\left(\bar{e}_{x_{(i)1}}-\bar{e}_{x_{(i)2}}\right)^2=\left(\theta_2-\theta_1\right)\bar{X}_i^2C^2_{x_i}\ \ etc\\
&E\left\{\bar{e}_{y_2}\left(\bar{e}_{x_{(i)1}}-\bar{e}_{x_{(i)2}}\right)\right\}=\left(\theta_1-\theta_2\right)\bar{X}_i\bar{Y}C_{x_i}C_y\rho_{yx_i}\ \ etc
\end{aligned}\right\}$$

(1.4.2)

The collection of correlation coefficients of various variables will be embedded in the following matrices:

$\mathbf{R}_{y\mathbf{x}}$: Correlation matrix for all variables under study

$\mathbf{R}_{\mathbf{x}}$: Correlation matrix for all auxiliary variables

$\mathbf{R}_{yx_i}$: Minor of $\mathbf{R}_{y\mathbf{x}}$ corresponding to ρ_{yx_i}

The following popular result of Arora and Lal (1989) will also be used:

$$\left|\mathrm{R}_{yx}\right|\big/\left|\mathrm{R}_x\right|=\left(1-\rho^2_{y.\mathrm{x}}\right). \qquad (1.4.3)$$

We will also use the vector notations in case of multiple auxiliary variables and the sample mean vector of auxiliary variables at *h*th phase will be denoted by $\bar{\mathrm{x}}_h$ with relation $\bar{\mathrm{x}}_h=\bar{\mathrm{X}}+\bar{\mathrm{e}}_{x_h}$. The following additional expectations are also useful:

$$E\left(\bar{\mathrm{e}}_{x_h}\bar{\mathrm{e}}^{/}_{x_h}\right)=\theta_h\mathrm{S}_x\ ,\ E\left(\bar{e}_{y_1}\bar{\mathrm{e}}_{x_h}\right)=\theta_1\mathrm{s}_{yx}\ ;h\geq 1 \qquad (1.4.4)$$

Similar expectations for qualitative auxiliary variables are:

$$\left.\begin{aligned}
&E\left(\bar{e}_{\tau_1}^2\right)=\theta_1 P_\tau^2 C_\tau^2\ ;\ E\left(\bar{e}_{\delta_1}^2\right)=\theta_1 P_\delta^2 C_\delta^2\ ;\ E\left(\bar{e}_{y_1}^2\right)=\theta_1 \bar{Y}^2 C_y^2\\
&E\left(\bar{e}_{\tau_2}^2\right)=\theta_2 P_\tau^2 C_\tau^2\ ;\ E\left(\bar{e}_{\delta_2}^2\right)=\theta_2 P_\delta^2 C_\delta^2\ ;\ E\left(\bar{e}_{y_2}^2\right)=\theta_2 \bar{Y}^2 C_y^2\\
&E\left(\bar{e}_{\tau_1}\bar{e}_{\tau_2}\right)=\theta_1 P_\tau^2 C_\tau^2\ ;\ E\left(\bar{e}_{\delta_1}\bar{e}_{\delta_2}\right)=\theta_1 P_\delta^2 C_\delta^2\\
&E\left(\bar{e}_{\tau_1}\bar{e}_{y_1}\right)=\theta_1\bar{Y}P_\tau C_\tau C_y\rho_{pb};E\left(\bar{e}_{\tau_1}\bar{e}_{y_2}\right)=E\left(\bar{e}_{\tau_2}\bar{e}_{y_1}\right)=\theta_1\bar{Y}P_\tau C_\tau C_y\rho_{pb}\\
&E\left(\bar{e}_{\tau_2}\bar{e}_{y_2}\right)=\theta_2\bar{Y}P_\tau C_\tau C_y\rho_{pb}\ ;\ E\left(\bar{e}_{\tau_1}\bar{e}_{\delta_1}\right)=\theta_1 P_\tau P_\delta C_\tau C_\delta Q_{xz}\ \ etc\\
&E\left(\bar{e}_{\tau_1}-\bar{e}_{\tau_2}\right)^2=\left(\theta_2-\theta_1\right)P_\tau^2 C_\tau^2\ \ etc\\
&E\left\{\bar{e}_{y_2}\left(\bar{e}_{\tau_1}-\bar{e}_{\tau_2}\right)\right\}=\left(\theta_1-\theta_2\right)\bar{Y}P_\tau C_\tau C_y\rho_{pb}\ \ etc
\end{aligned}\right\}$$

(1.4.5)

For multivariate case we will use following notations:

$$\left.\begin{aligned}
&E\left(\bar{\mathrm{e}}_{\mathrm{y}_1}\bar{\mathrm{e}}_{\mathrm{y}_1}^{/}\right)=\theta_1\mathrm{S}_{\mathrm{y}}\ ;\ E\left(\bar{\mathrm{e}}_{\mathrm{x}_1}\bar{\mathrm{e}}_{\mathrm{x}_1}^{/}\right)=\theta_1\mathrm{S}_{\mathrm{x}}\\
&E\left(\bar{\mathrm{e}}_{\mathrm{y}_2}\bar{\mathrm{e}}_{\mathrm{y}_2}^{/}\right)=\theta_2\mathrm{S}_{\mathrm{y}}\ ;\ E\left(\bar{\mathrm{e}}_{\mathrm{x}_2}\bar{\mathrm{e}}_{\mathrm{x}_2}^{/}\right)=\theta_2\mathrm{S}_{\mathrm{x}}\\
&E\left(\bar{\mathrm{e}}_{\mathrm{y}_1}\bar{\mathrm{e}}_{\mathrm{x}_1}^{/}\right)=\theta_1\mathrm{S}_{\mathrm{yx}}\ ;\ E\left(\bar{\mathrm{e}}_{\mathrm{x}_1}\bar{\mathrm{e}}_{\mathrm{y}_1}^{/}\right)=\theta_1\mathrm{S}_{\mathrm{xy}}\\
&E\left(\bar{\mathrm{e}}_{\mathrm{y}_2}\bar{\mathrm{e}}_{\mathrm{x}_2}^{/}\right)=\theta_2\mathrm{S}_{\mathrm{yx}}\ ;\ E\left(\bar{\mathrm{e}}_{\mathrm{x}_2}\bar{\mathrm{e}}_{\mathrm{y}_2}^{/}\right)=\theta_2\mathrm{S}_{\mathrm{xy}}\\
&E\left(\bar{\mathrm{e}}_{\mathrm{y}_1}\bar{\mathrm{e}}_{\mathrm{x}_2}^{/}\right)=\theta_1\mathrm{S}_{\mathrm{yx}}\ =E\left(\bar{\mathrm{e}}_{\mathrm{y}_2}\bar{\mathrm{e}}_{\mathrm{x}_1}^{/}\right)\\
&E\left(\bar{\mathrm{e}}_{\mathrm{x}_1}\bar{\mathrm{e}}_{\mathrm{y}_2}^{/}\right)=\theta_1\mathrm{S}_{\mathrm{xy}}\ =E\left(\bar{\mathrm{e}}_{\mathrm{x}_2}\bar{\mathrm{e}}_{\mathrm{y}_1}^{/}\right)
\end{aligned}\right\}\qquad (1.4.6)$$

where $\mathbf{S}_{\mathbf{y}}$ is covariance matrix of variables of interest and so on. We will use the notation $t_{1(1)}, t_{2(1)}, t_{3(1)},.....$ to denote single-phase sampling estimators whereas estimators for two phase sampling will be denoted by $t_{1(2)}, t_{2(2)}, t_{3(2)},.....$ etc. We will also use the notations $n,\theta,\bar{x},\bar{y}\ and\ \bar{z}$ etc. for sample size and mean(s) whenever only single-phase sampling is involved. The notations $n_1,\theta_1,\bar{x}_1,\bar{y}_1\ and\ \bar{z}_1$ will not be used in that case. We will also use $\bar{x}=\bar{X}+\bar{e}_x$ etc. to describe the relationship between sample mean and population mean of a variable whenever only single-phase sampling is involved.

1.5 Mean per Unit Estimator in Simple Random Sampling

The mean per unit estimator is perhaps the oldest estimator in the history of survey sampling. This estimator has provided the basis for the development of a number of single phase and two phase sampling estimators. The estimator for a sample of size *n* drawn from a population of size *N* is defined as:

$$t_0 = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}\,. \tag{1.5.1}$$

Using the notation defined in Section (1.4), i.e. $\bar{y} = \bar{Y} + \bar{e}_y$, we can write $t_0 - \bar{Y} = \bar{e}_y$. The mean square error (variance; as estimator is unbiased) can be immediately written as:

$$MSE\left(t_0\right) = \theta \bar{Y}^2 C_y^2\,. \tag{1.5.2}$$

The mean square error of t_0; given in (1.5.2) has been the basis in numerous comparative studies. The expression (1.5.2) has also been used by a number of survey statisticians to see the design effect of newly developed sampling designs.

Searle (1964) presented a modified version of mean per unit estimator as given below:

Suppose

$$t_s = k\bar{y}\,, \tag{1.5.3}$$

where k is a constant which is determined by minimizing mean square error of (1.5.3). The mean square error of (1.5.3) is readily written as:

$$MSE\left(t_s\right) = E\left(t_s - \bar{Y}\right)^2 = E\left(k\,\bar{y} - \bar{Y}\right)^2\ . \tag{1.5.4}$$

Expanding the square and rearranging the terms, the mean square error of t_s may be written as:

$$MSE\left(t_s\right) = E\left[k^2\left(\bar{y} - \bar{Y}\right)^2 + \left(k-1\right)^2 \bar{Y}^2 + 2k\left(k-1\right)\bar{Y}\left(\bar{y} - \bar{Y}\right)\right]$$

Applying expectation and simplifying we may write the mean square error as:

$$MSE\left(t_s\right) \approx \bar{Y}^2\left[\theta k^2 C_y^2 + \left(k-1\right)^2\right]. \tag{1.5.5}$$

Differentiating (1.5.5) *w.r.t. k* and equating to zero, we obtain optimum value of *k* which minimizes (1.5.4) as:

$$k=\left(1+\theta C_y^2\right)^{-1}. \tag{1.5.6}$$

Using *k* from (1.5.6) in (1.5.5) and simplifying, the mean square error of t_s is:

$$MSE\left(t_s\right)\approx\frac{\theta\bar{Y}^2 C_y^2}{1+\theta C_y^2}=\frac{MSE\left(t_0\right)}{1+\bar{Y}^{-2}MSE\left(t_0\right)} \tag{1.5.7}$$

Comparison of (1.5.7) with (1.5.2) shows that the estimator proposed by Searle (1964) is always more precise than the mean per unit estimator.

The estimator proposed by Searle (1964) is popularly known as Shrinkage mean per unit estimator. Many statisticians have studied the shrinkage version of popular survey sampling estimators. The focus of these studies had been reduction in mean square error of estimate or increase in precision. We will present some popular shrinkage estimators in this monograph. A general system of development of shrinkage estimators has been proposed by Shahbaz and Hanif (2009).

1.6 Shrinkage Estimator by Shahbaz and Hanif

Suppose a population parameter Θ can be estimated by using an estimator $\hat{\eta}$ with mean square error $MSE\left(\hat{\eta}\right)$. Shahbaz and Hanif (2009) defined a general shrinkage estimator as $\hat{\eta}_s=d\hat{\eta}$ where *d* is a constant to be determined so that mean square error of $\hat{\eta}_s$ is minimized. The optimum value of *d* and mean square error of $\hat{\eta}_s$ can be derived by assuming that $\hat{\eta}=\Theta+\bar{e}_\eta$ with $E\left(\bar{e}_\eta\right)=0$ and $E\left(\bar{e}_\eta^2\right)=MSE\left(\hat{\eta}\right)$. The mean square error of $\hat{\eta}_s$ and optimum value of *d*; under the above assumption is derived below:

The mean square error of $\hat{\eta}_s$ is given as:

$$MSE\left(\hat{\eta}_s\right)=E\left(\hat{\eta}_s-\Theta\right)^2=E\left(d\hat{\eta}-\Theta\right)^2=E\left[d\left(\Theta+\bar{e}_\eta\right)-\Theta\right]^2. \tag{1.6.1}$$

Differentiating (1.6.1) *w.r.t. d* and equating to zero we have:

$$E\left[\left\{d\left(\Theta+\bar{e}_\eta\right)-\Theta\right\}\left(\Theta+\bar{e}_\eta\right)\right]=0, \tag{1.6.2}$$

or $d\left[\Theta^2+E\left(\bar{e}_\eta^2\right)+2\Theta E\left(\bar{e}_\eta\right)\right]-\Theta^2-\Theta E\left(\bar{e}_\eta\right)=0$.

Now using the fact that $E\left(\bar{e}_\eta\right)=0$ and $E\left(\bar{e}_\eta^2\right)=MSE\left(\hat{\eta}\right)$, the above equation may be written as:

$$d\left[\Theta^2+MSE\left(\hat{\eta}\right)\right]-\Theta^2=0,$$

or $d=\Theta^2\left[\Theta^2+MSE\left(\hat{\eta}\right)\right]^{-1}=\left[1+\Theta^{-2}MSE\left(\hat{\eta}\right)\right]^{-1}$. (1.6.3)

Now from (1.6.1) we may write:

$$MSE\left(\hat{\eta}_s\right)=E\left[\left\{d\left(\Theta+\bar{e}_\eta\right)-\Theta\right\}\left\{d\left(\Theta+\bar{e}_\eta\right)-\Theta\right\}\right],$$

or $MSE\left(\hat{\eta}_s\right)=E\left[\left\{d\left(\Theta+\bar{e}_\eta\right)-\Theta\right\}d\left(\Theta+\bar{e}_\eta\right)\right]-\Theta E\left[\left\{d\left(\Theta+\bar{e}_\eta\right)-\Theta\right\}\right]$.

(1.6.4)

Also from (1.6.2) we have $E\left[\left\{d\left(\Theta+\bar{e}_\eta\right)-\Theta\right\}d\left(\Theta+\bar{e}_\eta\right)\right]=0$. Then (1.6.4) will be

$$MSE\left(\hat{\eta}_s\right)=-\Theta E\left[\left\{d\left(\Theta+\bar{e}_\eta\right)-\Theta\right\}\right]=-\Theta\left[dE\left(\bar{e}_\eta\right)-\Theta\left(1-d\right)\right],$$

or $MSE\left(\hat{\eta}_s\right)=\Theta^2\left(1-d\right)$ (1.6.4)

Using (1.6.3) in (1.6.4), the mean square error of $\hat{\eta}_s$ is:

$$MSE\left(\hat{\eta}_s\right)=\frac{MSE\left(\hat{\eta}\right)}{1+\Theta^{-2}MSE\left(\hat{\eta}\right)}. \quad (1.6.5)$$

The expression for mean square error given in (1.6.5) may be used to obtain the mean square error of the shrinkage version of any estimator. Also the shrinkage version of any estimator can be obtained by multiplying (1.6.3) with estimator to be shrunk. The estimator proposed by Searle (1964) turned out to be a special case of the shrinkage estimator proposed by Shahbaz and Hanif (2009) by using $\hat{\eta}=\bar{y}$.

An alternative derivation of the above estimator suggested by H.P. Singh (2011) is given as:

Consider

$$\hat{\eta}_s=d\hat{\eta}, \quad (1.6.6)$$

or $\left(\hat{\eta}_s-\Theta\right)=\left(d\hat{\eta}-\Theta\right)=\left\{d\left(\hat{\eta}-\Theta\right)+\left(d-1\right)\Theta\right\}$. (1.6.7)

Squaring both sides of (1.6.7) we have:

$$\left(\hat{\eta}_s-\Theta\right)^2=\left\{d\left(\hat{\eta}-\Theta\right)+\left(d-1\right)\Theta\right\}^2,$$

$$=d^2\left(\hat{\eta}-\Theta\right)^2+\left(d-1\right)^2\Theta^2+2d\left(d-1\right)\Theta\left(\hat{\eta}-\Theta\right).$$

Taking expectation of both sides of the above expression we get the MSE of $\hat{\eta}_s$ as

$$MSE\left(\hat{\eta}_s\right)=d^2MSE\left(\hat{\eta}\right)+\Theta^2\left(d-1\right)^2+2d\left(d-1\right)\Theta B\left(\hat{\eta}\right)$$

$$=\left[d^2\left\{\Theta^2+2\Theta B\left(\hat{\eta}\right)+MSE\left(\hat{\eta}\right)\right\}-2d\left\{\Theta^2+\Theta B\left(\hat{\eta}\right)\right\}+\Theta^2\right]. \tag{1.6.8}$$

Expression (1.6.8) is minimized when

$$d=d_{opt}=\frac{\left\{\Theta^2+\Theta B\left(\hat{\eta}\right))\right\}}{\left\{\Theta^2+2\Theta B\left(\hat{\eta}\right)+MSE\left(\hat{\eta}\right)\right\}}. \tag{1.6.9}$$

Thus the resulting minimum MSE of $\hat{\eta}_s$ is given by

$$MSE_{\min}\left(\hat{\eta}_s\right)=\left[\Theta^2-\frac{\left\{\Theta^2+\Theta B\left(\hat{\eta}\right)\right\}^2}{\left\{\Theta^2+2\Theta B\left(\hat{\eta}\right)+MSE\left(\hat{\eta}\right)\right\}}\right],$$

$$=\frac{\Theta^2\left[MSE\left(\hat{\eta}\right)-\left\{B\left(\hat{\eta}\right)\right\}^2\right]}{\left[\Theta^2+2\Theta B\left(\hat{\eta}\right)+MSE\left(\hat{\eta}\right)\right]}. \tag{1.6.10}$$

If we assume that $\hat{\eta}$ is an unbiased estimator (i.e. $Var\left(\hat{\eta}\right)=MSE\left(\hat{\eta}\right)$ then (1.6.10) reduces to

$$MSE_{\min}\left(\hat{\eta}_s\right)=\frac{\Theta^2MSE\left(\hat{\eta}\right)}{\Theta^2+MSE\left(\hat{\eta}\right)}=\frac{MSE(\hat{\eta})}{1+\Theta^{-2}MSE\left(\hat{\eta}\right)}, \tag{1.6.11}$$

which is identical to (1.6.5) suggested by Shahbaz and Hanif (2009).

SECTION-I:

SINGLE-PHASE SAMPLING WITH AUXILIARY VARIABLE

CHAPTER TWO

SINGLE-PHASE SAMPLING USING ONE AND TWO AUXILIARY VARIABLES

2.1 Introduction

Survey statisticians are always trying to estimate population characteristics with greater precision. The precision of estimate can be increased by using two methodologies. Firstly the precision may be increased by using adequate sampling design for the estimated variable (estimand). Secondly the precision may be increased by using an appropriate estimation procedure, i.e. some auxiliary information which is closely associated with the variable under study.

The ratio and regression methods of estimation are two classical methods of estimation in survey sampling which use information about auxiliary variables. These two methods of estimation are more efficient than the mean per unit method of estimation in terms of smaller mean square error. Various estimators have been proposed in the literature that use information on multiple auxiliary variables and consequently produce smaller error as compared with estimators based upon single auxiliary variables.

Various authors have proposed mixed type estimators, i.e. use of both ratio and regression estimators in some fashion. These mixed estimators have been seen performing better as compared with individual estimators. Mohanty (1967) used this methodology for the first time to propose mixed estimator using two auxiliary variables.

In this Chapter we will explore some popular single-phase sampling estimators which utilize information of one and two auxiliary variables. We will first reproduce some estimators which are based upon single auxiliary variable and then we will describe those estimators which are based upon two auxiliary variables.

2.2 Estimators using Single Auxiliary Variable

Suppose that information on a single auxiliary X is available for complete population units from some previous source. This available information of X may be very useful in increasing the precision of estimate. In the following we describe some popular estimators which are based upon information of a single auxiliary variable.

2.2.1 Classical Ratio Estimator

The classical ratio estimator is perhaps the most celebrated estimator in single phase sampling. The estimator based upon single auxiliary variable is given as:

$$t_{1(1)} = \bar{y}\frac{\bar{X}}{\bar{x}} = \frac{\bar{y}}{\bar{x}}\bar{X} \; . \qquad (2.2.1)$$

Using the notations from section (1.4) we may write the estimator (2.2.1) as:

$$t_{1(1)} = \left(\bar{Y} + \bar{e}_y\right)\frac{\bar{X}}{\bar{X} + \bar{e}_x} = \left(\bar{Y} + \bar{e}_y\right)\left(1 + \frac{\bar{e}_x}{\bar{X}}\right)^{-1} .$$

Expanding $\left(1 + \bar{e}_x / \bar{X}\right)^{-1}$, under the assumption that $\left|\bar{e}_x / \bar{X}\right| < 1$, and retaining the first term only and by using argument of Cochran (1940); we have:

$$t_{1(1)} = \left(\bar{Y} + \bar{e}_y\right)\left(1 - \frac{\bar{e}_x}{\bar{X}}\right) = \bar{Y} + \bar{e}_y - \frac{\bar{Y}}{\bar{X}}\bar{e}_x \, .$$

The mean square error of $t_{1(1)}$ is

$$\mathrm{MSE}\left(t_{1(1)}\right) = E\left(t_{1(1)} - \bar{Y}\right)^2 = E\left(\bar{e}_y - \frac{\bar{Y}}{\bar{X}}\bar{e}_x\right)^2 \cdot \qquad (2.2.2)$$

Squaring, applying expectation and simplifying we get

$$MSE\left(t_{1(1)}\right) = \theta\bar{Y}^2\left(C_y^2 - 2C_xC_y\rho_{xy} + C_x^2\right),$$

or $$MSE\left(t_{1(1)}\right) = \theta\bar{Y}^2\left[C_y^2 - 2C_xC_y\rho_{xy} + C_x^2 + C_y^2\rho_{xy}^2 - C_y^2\rho_{xy}^2\right],$$

or $$MSE\left(t_{1(1)}\right) = \theta\bar{Y}^2\left[C_y^2 - 2C_xC_y\rho_{xy} + C_x^2 + C_y^2\rho_{xy}^2 - C_y^2\rho_{xy}^2\right]\cdot \qquad (2.2.3)$$

Comparing (2.2.3) with variance of mean per unit estimator we can see that the ratio estimator will have larger precision as compared with mean per unit estimator if

$$\rho_{xy} > C_x \big/ 2C_y \cdot \qquad (2.2.4)$$

For a smaller correlation coefficient, the mean per unit estimator will be more precise. Survey statisticians have proposed some modifications of classical ratio estimator from time to time. In the following we discuss some of the popular modifications of the classical ratio estimator.

2.2.2 Prasad's (1989) Ratio Estimator

Prasad (1989) proposed following modification of ratio estimator following the lines of Searle (1964):

$$t_{2(1)} = \frac{k\bar{y}}{\bar{x}}\bar{X} \,, \qquad (2.2.5)$$

where k is constant. Using Section (2.5), $\beta_i = -\left(1-2k_i\right)\bar{Y} \big/ \bar{X}_i$ may be written as

$$t_{2(1)} = k\bar{Y}\left(1+\frac{\bar{e}_y}{\bar{Y}}\right)\left(1+\frac{\bar{e}_x}{\bar{X}}\right)^{-1} ,$$

or $$t_{2(1)} = k\bar{Y}\left[1+\frac{\bar{e}_y}{\bar{Y}}-\frac{\bar{e}_x}{\bar{X}}-\frac{\bar{e}_y\bar{e}_x}{\bar{X}\,\bar{Y}}+\frac{\bar{e}_x^2}{\bar{X}^2}\right],$$

or $$t_{2(1)} - \bar{Y} = \bar{Y}\left[k\left(1+\frac{\bar{e}_y}{\bar{Y}}-\frac{\bar{e}_x}{\bar{X}}-\frac{\bar{e}_y\bar{e}_x}{\bar{X}\,\bar{Y}}+\frac{\bar{e}_x^2}{\bar{X}^2}\right)-1\right].$$

Squaring the above equation and ignoring higher order terms we have:

$$\left(t_{2(1)}-\bar{Y}\right)^2 = \bar{Y}^2\left[k^2\left\{1+\frac{\bar{e}_y^2}{\bar{Y}^2}+\frac{3\bar{e}_x^2}{\bar{X}^2}+\frac{2\bar{e}_y}{\bar{Y}}-\frac{2\bar{e}_x}{\bar{X}}-\frac{4\bar{e}_x\bar{e}_y}{\bar{Y}\,\bar{X}}\right\}\right.$$
$$\left. +1-2k\left\{1+\frac{\bar{e}_y}{\bar{Y}}-\frac{\bar{e}_x}{\bar{X}}-\frac{\bar{e}_y\bar{e}_x}{\bar{X}\bar{Y}}+\frac{\bar{e}_x^2}{\bar{X}^2}\right\}\right].$$

Applying expectations on the above equation, the mean square error of $t_{2(1)}$, to first degree of approximation is

$$MSE\left(t_{2(1)}\right) = \bar{Y}^2\left[k^2\left\{1+\theta(C_y^2+3C_x^2-4\rho_{xy}C_xC_y)\right\}\right.$$

$$+1-2k\left\{1+\theta(C_x^2-\rho_{xy}C_xC_y)\right\}\Big], \tag{2.2.6}$$

which is minimum when

$$k=k_{opt}=\frac{\left\{1+\theta\left(C_x^2-\rho_{xy}C_xC_y\right)\right\}}{\left\{1+\theta\left(C_y^2+3C_x^2-4\rho_{xy}C_xC_y\right)\right\}}.$$

Using k_{opt} in (2.2.6) and simplifying, the mean square of $t_{2(1)}$ is:

$$MSE_{\min}\left(t_{2(1)}\right)=\bar{Y}^2\left[1-\frac{\left\{1+\theta\left(C_x^2-\rho_{xy}C_xC_y\right)\right\}^2}{\left\{1+\theta\left(C_y^2+3C_x^2-4\rho_{xy}C_xC_y\right)\right\}}\right],$$

or $$MSE\left(t_{2(1)}\right)=\frac{\left[MSE\left(t_{1(1)}\right)-\left\{B\left(t_{1(1))}\right)\right\}^2\right]}{\left[1+RMSE\left(t_{1(1)}\right)+2RB\left(t_{1(1)}\right)\right]}, \tag{2.2.7}$$

where $RB\left(t_{1(1)}\right)=\bar{Y}^{-1}B\left(t_{1(1)}\right)$ and $RMSE\left(t_{1(1)}\right)=\bar{Y}^{-2}MSE\left(t_{1(1)}\right)$.

The above result can be easily obtained from the general result given in Section 1.6.

2.2.3 Product Estimator

The ratio estimator and its versions, which we will discuss later, are useful for estimating population mean and total when the auxiliary variable has a high positive correlation with the variable of interest. Sometimes the available auxiliary variable has a negative correlation with the variable of interest; the survey statisticians have proposed alternative methods of estimation in such situations. One such method of estimation is the product method which was initially studied by Robson (1957) and Murthy (1964). Cochran (1977), Chaubey et al. (1984), and Kaur (1984) have also done substantial work on it. The conventional product estimator is

$$t_{3(1)}=\bar{y}\frac{\bar{x}}{\bar{X}}. \tag{2.2.8}$$

Using (1.4) in (2.2.8) and on simplification we get: