

**RATIO ESTIMATION OF FINITE
POPULATION TOTAL IN STRATIFIED
RANDOM SAMPLING UNDER
NON-RESPONSE**

A Thesis Submitted to the University of Nairobi for the Award
of the Degree of Doctor of Philosophy in Mathematical Statistics
in the School of Mathematics

OYOO DAVID ODHIAMBO

I80/50476/2016

2021

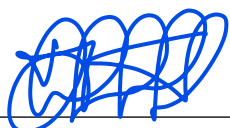
Declaration

This thesis is my original work and has not been presented for award of a degree in any other university.

David Odhiambo Oyoo

I80/50476/2016

Signed:



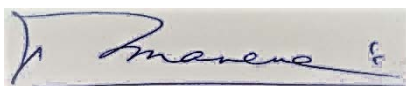
Date:

30/11/2021

This work has been submitted with our approval as the university supervisors.

Prof. Mosses M. Manene

Signed:

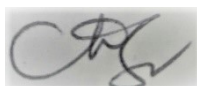


Date:

01/12/2021

Dr. Christopher O. Ouma

Signed:



Date:

01/12/2021

Dr. George Muhua

Signed:



Date:

1/12/2021

Dedication

To my wife Asha, daughter Lexi, son Dalex, my parents and siblings

Acknowledgment

First and foremost my gratitude goes to God Almighty, who without His grace and power, this work would not have been possible.

My special gratitude to my supervisors Prof. Moses Manene, Dr. Christopher Ouma and Dr. George Muhua for providing invaluable scholarly comments, guidance and support that greatly shaped my research work, and for creating time to read this work at different stages despite their busy schedules. A special gratitude goes to the School Directors and Heads of Statistics Departments both at the University of Nairobi and at Technical University of Kenya.

I am also grateful to the University of Nairobi for according me the opportunity to do my PhD and giving me the support needed to complete the study. I appreciate the support and love from all my colleagues from the School of Mathematics and all my friends who have not been mentioned by name for assisting me in various ways.

I am greatly indebted to all members of my family, my brothers and sisters who were a source of inspiration. Kindly receive my appreciation for supporting me in all aspects of life and inculcating the love for education in me. Lastly, a special gratitude to my lovely wife for her concern and for always being there for me.

To anyone who made this study possible, I cannot list you all. I highly appreciate your positive criticisms during the seminars and presentation meetings. Your input to my thesis was indispensable. I, from the bottom of my heart, appreciate your role.

Abstract

Estimation of population parameters has been an area of interest to many statisticians. Auxiliary variable that is highly correlated with the response variable can be used to improve efficiency of constructed estimators. Efficiency of constructed estimators is improved when more auxiliary variables are used in the survey problems. However, asymptotic properties of constructed estimators are usually interfered with by non-response in the study variable. Various corrective measures, such as imputation, partial deletion, resampling, weight adjustment and sub-sampling, have been suggested in literature to take care of the non-response. In this study, we have adopted the sub-sampling approach to construct a ratio estimator for finite population total in stratified random sampling under non-response. This has been done under both univariate and multivariate ratio estimations. In univariate case, we have considered separate and combined ratio estimations and regression forms of the constructed estimator. From the Percent Relative Efficiency (PRE) computations, we have observed that stratification improves performance of the constructed estimator by 10.26% compared to simple random sampling without replacement. Also, the sub-sampling method adopted improved efficiency of the constructed estimator by 0.44% when partial deletion is used. From multivariate unbiased ratio estimation, a two dimensional auxiliary random vector was constructed and it was observed that performance of the constructed multivariate ratio estimators depends on the choice of multivariate weights. This study has shown how an unbiased ratio estimator for finite population total is constructed in stratified random sampling. The study has also shown how the problem of non-response in sample surveys can be corrected using sub-sampling method.

CONTENTS

<i>Declaration</i>	ii
<i>Dedication</i>	iii
<i>Acknowledgment</i>	iv
<i>Abstract</i>	v
<i>Terminologies and Definitions</i>	viii
<i>Acronyms</i>	x
<i>List of Tables</i>	xi
1. INTRODUCTION	1
1.1 Introduction	1
1.2 Background of the Study	1
1.3 Statement of Problem	5
1.4 Objectives of the Study	6
1.5 Significance of the Study	7
1.6 Areas of Application	7
1.7 Assumptions in the Study	8
2. LITERATURE REVIEW	9
2.1 Introduction	9
2.2 Ratio Estimation Method	9
2.3 Non-Response in Sample Surveys	19
2.4 Summary of Study Gaps	24

3. <i>METHODOLOGY</i>	25
3.1 Introduction	25
3.2 Notation and Definition of Symbols	25
3.3 Ratio Estimation of Population Total in Stratified Random Sampling	27
3.4 Regression Estimation	30
3.5 Multivariate Unbiased Ratio Estimation	31
3.6 Weaknesses of Reviewed Estimators	32
3.7 Construction of Improved Estimator	33
3.8 Simulation Study	34
4. <i>UNBIASED RATIO ESTIMATOR</i>	36
4.1 Introduction	36
4.2 Unbiased Separate Ratio-Type Estimator	36
4.3 Combined Ratio Form of \hat{Y}_D	56
4.4 Comparison of Separate and Combined Ratio Estimators	65
4.5 Chapter Summary	67
5. <i>REGRESSION-BASED UNBIASED RATIO ESTIMATORS</i>	68
5.1 Introduction	68
5.2 Regression Form of \hat{Y}_D	68
5.3 Efficiency of \hat{Y}_{DR}	77
5.4 Multivariate Form of \hat{Y}_D	79
5.5 Simulation Study	89
5.6 Chapter Summary	102
6. <i>SUMMARY, CONCLUSIONS AND RECOMMENDATIONS</i>	104
6.1 Introduction	104
6.2 Summary	104
6.3 Conclusion	106
6.4 Contributions of the Study to Knowledge	107
6.5 Recommendations for Further Research	108
BIBLIOGRAPHY	109
<i>Appendix A. Simulation Code</i>	115

Terminologies and Definitions

Auxiliary Variable: It is the additional information available on every population unit apart from the information (variable) of interest. Auxiliary information is not only used to improve sampling plan, but also to enhance estimation of the variables of interest.

Non-Response: In surveys, non-response occurs when there is a failure to make observation on or obtain data for some population units.

Non-Response Bias: Is a bias that occurs in surveys when observations cannot be made on some population units due to some factors that make them differ significantly from units whose observations can be made.

Optimality Condition: An estimator $\hat{\theta}^*$ is said to satisfy optimality conditions in estimating an unknown population parameter θ (say), if it is a best linear unbiased estimator (BLUE) among a class of unbiased estimators $\hat{\theta}$ of θ .

Population: Set of individuals, items or objects that share or have at least one observable characteristics in common that can be studied.

Population Periodicity: It is the frequency of any observed pattern in the population especially after identifying and ordering population units. The pattern can be based on some auxiliary variable(s) or attribute possessed by population units.

Proportional Allocation: Is allocation of sample sizes in a stratified population such that in each stratum, the sampling fraction n/N remains constant.

Response Variable: It is the variable under study or the variable of interest in surveys.

Statistical Sample: Is a representative portion or subset of a population and is used for statistical analysis.

Unit: The element of analysis.

Acronyms

COD: Coefficient of determination

GDP: Gross Domestic Product

MVUE: Minimum variance unbiased estimator

PRE: Percent relative efficiency

SRSWR: Simple random sampling with replacement

SRSWOR: Simple random sampling without replacement

LIST OF TABLES

Table 5.1	Stratum Population and Sample Sizes	89
Table 5.2	Random Data Parameters	90
Table 5.3	Summary Statistics for SRSWOR	92
Table 5.4	Summary Statistics	93
Table 5.5	Results of Univariate Estimation	94
Table 5.6	Summary Statistics for X_2	97
Table 5.7	Results for Multivariate Estimation	99
Table 5.8	Univariate Variances and Optimal Weights	101
Table 5.9	Minimum Variance under Optimal Weights	102

1. INTRODUCTION

1.1 Introduction

This chapter presents background of the study, statement of the problem and specific objectives of the study. The chapter also discussed significance of the study and main assumptions in the study.

1.2 Background of the Study

Sampling involves selection of a representative subset of individuals or units from a population. A statistical sample is therefore a data set collected from a population using some defined procedure. Elements of a statistical sample are referred to as sample units or sample points or generally, observations. Sample observations are used to estimate and make inferences about the population characteristics. Sampling is, therefore, the process of choosing sample units from a population (Murthy, 1967). Some advantages of survey samples over complete enumeration include reduced cost and time of conducting a survey and improved accuracy among others (Dorofeev & Grant, 2006).

Even though sample surveys, under certain conditions, can be preferred to complete enumeration, its effectiveness, however, depends on how sample units are selected, how observation of sampled units is made and how inferences about population parameters is made using sample observations. In particular, the questions have been on how observations should be made, the number of observations to be used, how to analyze the obtained data and how to interpret and make inferences using values from the analysis (Singh & Mangat, 1996). Solutions to these inquiries have led to coming up with different techniques and methods in survey sampling. Sample survey theory is, therefore, concerned with development of sampling procedures that

yield a sample which best represents the entire population (Sampath, 2001).

Sampling strategies can be developed using three main approaches; model-based, model-assisted and design-based approaches (Pfefferman & Rao, 2009). Model based methods, in estimation of population parameters, assume that there is an underlying probability model that generate survey units. However, if the assumed model is incorrect, estimators of population parameters will certainly be incorrect. In model-assisted estimation, non-sampled units are assumed to be unknown and therefore, the known sampled units of the population are used to estimate the unknown non-sampled units. This can be done using a suitable model or by a non-parametric approach. Use of regression models is an example of model-assisted estimation of population parameters. Design-based estimations, on the other hand, involve use of a probability mechanism to select a sample. This estimation procedure assumes a known sampling design and that the sample is large.

A sample can be selected using either subjective (non-probability) methods or probability sampling methods (Dorofeev & Grant, 2006). In subjective method, specific population units that bear specific traits, according to the researcher, are sampled. Examples of subjective sampling include convenience or accidental sampling, quota sampling and snow ball sampling. In convenience or accidental sampling, sample units are selected as they become available to the researcher. For instance, asking questions to radio listeners is a case of accidental or convenience sampling. In quota sampling, selection of sample units is done to fit some pre-identified quotas such as religion, academic levels or socio-economic class among others. For snowball sampling, referral networks is used since the researcher does not know specific population units possessing the characteristic under study.

Probability sampling methods involve selection procedures where each population unit has some assigned probability of inclusion in the sample. These sampling methods include simple random sampling (SRS), stratified random sampling, cluster sampling, systematic sampling and double and multistage sampling techniques. For SRS, all population units have equal and independent chance of being selected and included in the sample. In systematic sampling, some pre-designed intrinsic order is used to select sample units. In this case, all population units are identified and ordered either al-

phabetically or numerically and from among the first k^{th} units, a unit is selected at random. Suppose the u^{th} is selected, then the units included in the sample are $u, u + k, u + 2k, \dots, u + (n - 1)k$, that is every k^{th} after the first selected unit and any other selected unit. This constant u is called *random start* and n is the required sample size. Stratified random sampling, on the other hand, involves sub-dividing heterogeneous population into finite sub-populations known as strata such that there is a low intra-stratum variations but a high inter-stratum variations. A simple random sample is then obtained from each stratum and its observations are then used in estimation.

For cluster sampling, population units present themselves in finite groups of elements known as clusters such that instead of selecting individual population elements, the clusters are randomly selected. In a case where only a part of elements in a cluster are used in estimation, the resulting design is double-stage sampling method. Repeated sub-cluster units' selection, while narrowing the number of sampled units, results to multi-stage cluster sampling.

Each of these probability sampling methods have individual strengths and weaknesses. However, irrespective of the sampling method adopted in a research work, development of efficient estimators with minimum error is still a problem in sample surveys (Oyoo, Manene, Ouma & Muhua, 2019). To this end, studies focusing on error minimization while maximizing efficiency of constructed estimators have been conducted by different researchers. As a solution to this problem, several studies have applauded the use of population characteristic that is closely related to the study variable to improve efficiency of constructed estimators. Such population characteristic is called auxiliary characteristic. Use of auxiliary variable or attribute has led to evolution of ratio estimation in sample surveys. In ratio estimation, therefore, for any sampled unit, an observation can simultaneously be made for both the study variable (or attribute) and auxiliary variable (or attribute).

Ratio estimation improves efficiency of estimators of population parameters compared to simple mean per unit estimators (Cochran, 1977). The use of auxiliary variable to improve efficiency of estimators date back to mid 1600s when John Graunt estimated the total population size of England us-

ing some records of registered births in the preceding year (Pearson, 1897). By 1770s, the use of auxiliary variable in estimation became widespread in France. In 1802, Laplace successfully estimated the population of France in a population census using ratio estimation method. Karl Pearson, however, warned in 1897 that ratio estimator should be used with caution since their estimates are prone to bias (Pearson, 1897).

As noted by Swain (2014), Neyman used an auxiliary variable for stratification of a finite population. Later, Cochran (1940) used auxiliary information in estimation procedure and proposed ratio method of estimation to provide more efficient estimator of the population mean or total compared to the simple mean per unit estimator under the condition that the auxiliary variable has a strong correlation with the study variable. Auxiliary variable(s), therefore, play important role of reducing mean standard error of the estimates. Based on this observed role of auxiliary information, many authors have suggested estimators of population parameters using auxiliary variable(s). For instance, Kushwaha and Singh (1989) suggested a class of almost unbiased ratio and product type estimators for estimating the population mean using jack-knife technique initiated by Quenouille (1956). Afterward Banarasi et al. (1993) and Singh and Singh (1998) proposed the estimators of population mean using auxiliary information in systematic sampling.

Performance of any constructed estimator is significantly being influenced by whether there is non-response or not. Non-response results to missing value for a particular sampled unit and consequently, it distorts properties of estimators of finite population parameters. Various methods that can be used to address the problem of non-response in surveys include imputation, partial deletion, resampling, weight adjustment and sub-sampling among others (Daroga and Chaudhary, 2002). This study focuses on sub-sampling method suggested by Hansen and Hurwitz (1946). This method involves obtaining a random sample from a finite population and partitioning the sample into responding and non-responding groups. From the non-responding group, a sub-sample is made and an assumption that there will be total response in this sub-sample. This method improves efficiency of constructed estimators compared to other methods of addressing non-response.

1.3 *Statement of Problem*

Estimation of finite population parameters in sample surveys has been an area of concern in the recent past. Estimation of vectors of finite population parameters in small area estimation with or without auxiliary information has been considered by several authors when there is no measurement error. In a case where there are few observations on the response variable, sample survey estimators of finite population parameters will have large standard error. Based on this limitation, it is necessary to use information from similar neighboring or related variables for improved estimation. One way through which information from related variables can be utilized is through ratio estimation.

Using auxiliary variable, various ratio-type estimators have been constructed. Most of these estimators have, however, been constructed using the usual (biased) ratio. Despite this progress, non-response still remains to be a challenge in survey sampling. Non-respondents tend to have different attitudes towards survey questions, a situation that leads to non-response bias, which, consequently, influences inferences made about population parameters. This study is, therefore, motivated by this effect of non-response in survey sampling and has consequently considered constructing a ratio estimator for finite population total under non-response.

Apart from the problem of non-response, asymptotic properties of constructed estimators are affected by the type of sampling technique involved in a survey. Sampling technique adopted in a survey is dictated by the nature of survey population. Homogeneity of population units is one aspect of survey population that usually dictates sampling technique to be adopted while selecting sample units. Homogeneity of population units is not always guaranteed in surveys. Heterogeneity in a study population leads to construction of estimators with high standard errors. Use of stratified random sampling technique has, thus, been motivated by the fact that use of other sampling techniques in heterogeneous population yields estimators with high standard errors. That is, stratification enables construction of estimators with high precision. For this reason, this study involves construction of a ratio estimator for finite population total in stratified random sampling under non-response. Ratio estimation assumes a perfect linear relationship

between the response variable and auxiliary variable(s), which is not always the case. For this reason, this study considers regression form of the constructed ratio estimator to address non-perfect relationship.

Nature of constructed ratio estimators also depends on the dimensionality of study variables, especially when more than one auxiliary variable is involved. In cases where simultaneous observations and analysis of more than one study variables is made, the data set changes from uni-dimensional to multi-dimensional. This study considers multivariate estimation since in survey sampling, the response variable can be a function of more than one auxiliary variable, leading to complex data sets. In this study, therefore, we construct an unbiased ratio estimator for finite population total in stratified random sampling under non-response. We further consider the multivariate and regression forms of the constructed estimator.

1.4 Objectives of the Study

General Objective

To construct an unbiased ratio estimator for finite population total in stratified random sampling under non-response.

Specific Objectives

1. To construct an unbiased ratio estimator in stratified random sampling under non-response
2. To derive the regression form of the constructed unbiased ratio estimator
3. To derive the multivariate form of the constructed unbiased ratio estimator
4. To use simulated data to compare performance of the constructed unbiased ratio estimators

1.5 Significance of the Study

This study is significant in both scientific and societal development. In scientific development, the study is key in literature development since it has explained how an unbiased ratio estimator for finite population total in stratified random sampling is constructed under non-response. The study has also shown that sub-sampling method suggested by Hansen and Hurwitz (1946) to correct non-response produces efficient estimators compared to partial deletion, which is the commonly used corrective method. The study has also shown how the ratio estimator can be constructed when the response variable is a linear function of more than one auxiliary variable. The study has also confirmed a known knowledge that efficiency of ratio estimators is improved if the correlation between the response variable and auxiliary variable(s) is close to unity. Moreover, this study has demonstrated how to improve precision of ratio estimators for population parameters under non-response by stratifying the study population.

In societal development, surveys involving estimation of population totals or averages of various population characteristics is common in real life. However, such surveys are often accompanied by incomplete data, high defaulting rates or low response rates, which affects accuracy of observations made and bias estimates constructed. By having accurate information about a population and making correct inferences, appropriate measures are put in place to address a given problem within the survey population in question. This is only possible if a suitable mathematical model or estimator that is not only unbiased and efficient but also addresses the problem non-response is used in surveys.

1.6 Areas of Application

This study can effectively be used in socio-economic surveys where the focus is household ratios such as per household ratio of expenditure on various items, per capita income or expenditure or ratios of unemployed individuals. Similar application can also be done on epidemiological and demographic studies. By estimating population sizes of a country relative to its GDP, procedures discussed in this research work can help the government of Kenya in ensuring equitable distribution of available resources in

the entire country.

Similarly, in industrial surveys, this study can provide vital insights in studies involving input-output ratios. In agricultural sector, estimation and forecasting of agricultural produce is an area where this research work can extensively be applied. For banking and insurance sectors, estimation and forecasting of uptake of a given policy or service based on ratios of some characteristics of targeted consumers can easily be done by considering findings in this study. Other areas where findings of this study can effectively be applied include, among others, estimation of inflow and outflow of tourists in the tourism sector and estimation of hotel occupancy levels at various times of the year in the hotel industry. By correct adoption of the findings of this study in the above-stated sectors, this research work becomes integral in helping the government of Kenya to attain its vision 2030.

Also, this study has greatly improved the literature work on ratio estimation. By considering separate and combined ratio estimation methods under incomplete data, this study has significantly contributed to unbiased ratio estimation in stratified random sampling. Moreover, significant inputs in literature development have also been seen in the construction of multivariate and regression forms of the unbiased ratio estimator. The improvement is due to the fact that the constructed unbiased ratio-type estimators perform better than estimators in literature.

1.7 Assumptions in the Study

In this study, we have made the following assumptions:

- i) That the population size is large and correspondingly, a large sample size is randomly obtained
- ii) That there is a strong correlation between the study variable and auxiliary variable
- iii) That the auxiliary variable is independent of non-response

2. LITERATURE REVIEW

2.1 *Introduction*

This chapter reviews literature on the use of ratio estimation technique in stratified random sampling method and the concept of non-response in sample surveys.

2.2 *Ratio Estimation Method*

Statistical estimations aim at obtaining estimators for population parameters with high precision. This can be done by properly using any available auxiliary information through ratio method of estimation. In ratio estimation, some known information about the auxiliary variable is used to improve efficiency of constructed ratio-type estimators. To do this, ratio estimation entails adjusting the sample estimate of the study variable using the ratio of population mean (or total) of the auxiliary variable and the corresponding sample mean per unit estimate. For simplicity, Y and X shall be used to denote the response variable and auxiliary variable respectively. Ratio estimation is based on the assumption of existence of a linear relationship between X and Y (Murthy, 1967; Cochran, 1977; Daroga & Chaudhary, 2002). Utilization of auxiliary variable has led to construction of three main estimators, which are the traditional (usual) ratio estimator, product estimator and regression estimator. In this study, the focus is on the usual ratio estimator and regression estimator.

The usual ratio estimator has been shown to produce biased results irrespective of the nature of the linear relationship between X and Y (Cochran, 1977; Kadilar & Cingi, 2004; Singh & Smarandache, 2013). Despite this weakness, ratio estimator is, however, preferred to mean per unit estimator since it has a small variance compared to the counterpart (Cochran, 1977;

Daroga & Chaudhary, 2002). The issue of sample size notwithstanding, the main problem of ratio estimation has been how to reduce bias of a ratio estimator while upholding its good property of small variance. No conclusive method has been agreed upon on how to eliminate the bias or when to consider the bias negligible. Consequently, only bias approximations and limits for the bias at different orders have been suggested in literature (Zaman & Yilmaz, 2017). This problem has called for investigating ways of reducing bias of a ratio estimator and thus, constructing an unbiased ratio estimator. This can be done by either using the common sampling schemes or by modifying the common sampling schemes so that the usual (biased) ratio estimator becomes unbiased.

To maintain the property of minimum variance while estimating bias of ratio estimator, Cochran (1977) examined the conditions under which a ratio estimator is MVUE. For the first condition, he investigated the nature of the regression line of Y on X and he observed that if the line is straight and passes through the origin such that variance of Y is proportional to X about this line, then the ratio estimator becomes almost unbiased. Also, using Gauss-Markov Theorem, Cochran (1977) observed that for large sample, the distribution of the usual ratio estimator tends to normal distribution and that since bias of ratio estimator is of order $1/n$, then under these conditions, the ratio estimator becomes unbiased.

In the second condition, Cochran (1977) examined the coefficient of variation between X and Y and observed that for a large sample and if the correlation coefficient between X and Y is larger than half coefficient of variation of X divided by coefficient of variation of Y , variance of the ratio estimator becomes smaller than that of the unbiased mean per unit estimator.

Since the discovery of traditional ratio estimator, studies have been conducted with each suggesting how to best eliminate bias of the estimator. For instance, Koop (1951) used binomial series expansion of ratio estimator using various sample sizes in an effort to reduce bias of the ratio estimator to a desirable degree. Quennouille (1956), on the other hand, considered t_n , which is a function of sample observations, as a ratio estimator for an unknown population parameter. Using different sets of sample values through

repeated sampling using SRSWR and assuming that the estimator is consistent, Quennouille (1956) used Taylor series expansion and observed that the estimator becomes unbiased to order $1/n^2$.

While extending the work of Quennouille (1956), Durbin (1959) examined whether existence of linear relationship between X and Y has an effect on variance of the ratio estimator, which was confirmed to be true. By assuming that X has a gamma distribution, Durbin (1959) observed that existence of a linear relationship between X and Y significantly reduces MSE of the ratio estimator. However, as observed by Cochran (1977), Quennouille-based ratio estimators are more appropriate in large samples, which is not always the case. Other studies that have also suggested solutions to the problem of bias of ratio estimator during the early stages of ratio estimation included, among others, Hartley and Ross (1954), Jones (1956), Mickey (1958), Murthey and Nanjama (1960), Williams (1961) and Beale (1962).

Apart from these early work on ratio estimation that still did not fully eliminate bias of ratio estimator, Hartley and Ross (1954) were the first authors to consider common sampling techniques to construct unbiased ratio estimators. In their approach, Hartley and Ross (1954) evaluated bias of the ratio estimator and connected it to covariance of Y/X and X . Using this approach, an unbiased ratio estimator was constructed and its large sample variance was obtained. This unbiased ratio estimator was later studied by Goodman and Hartley (1958) for a sample of any size. Robson (1957) applied multivariate polykays on Hartley and Ross (1954) unbiased estimator to obtain exact variance of the estimator. Multivariate polykays, also known as multivariate generalized k-statistics and minimum variance unbiased estimators of joint cumulant products. Such estimators are often expressed in terms of power sum symmetric polynomial in the random vector of a sample Robson (1957). Some other studies that have considered common sampling schemes while constructing unbiased ratio estimators include Mickey (1958) and Williams (1961).

Using the second approach of modifying the sampling scheme, Lahiri (1951) showed that if sampling is done using probability proportional to sum of observations of auxiliary variable, the ratio estimator becomes unbiased. Lahiri (1951) further outlined the sampling procedure that yields this result.

The procedure outlined by Lahiri (1951) was initially studied and adopted by Nanjama, Murthey and Sethi (1960), Midzuno (1962) and Raj (1965).

Use of auxiliary information in survey sampling is not only limited to one auxiliary variable, but also more than one auxiliary variable. Several studies have considered more than one auxiliary variables or attributes in their estimation procedures. Some of these studies include Kadilar and Koyuncu (2009), Rao (2015), Kadilar and Cingi (2004), Kiregyera (1980) and Kiregyera (1984). With the use of more than one auxiliary variables, ratio estimation procedures have been classified into either univariate or multivariate. In univariate ratio estimation, the response variable is considered as a linear function of a one-component auxiliary random vector (Olkin, 1958). These mentioned studies that have considered more than one auxiliary variable were univariate in nature. In multivariate case, the auxiliary information is a p -dimensional random vector such that we have X_1, X_2, \dots, X_p . In this case, the parameter to be estimated is assumed to be a function of these p auxiliary variables (Olkin, 1958).

Use of multi-auxiliary variables to construct a multivariate ratio estimator for population parameters date back to 1958 when Olkin (1958) suggested a multivariate ratio estimator for population total under simple random sampling scheme. Since then, improvements on Olkin's estimator have been done to reduce its bias and MSE. Also, other sampling schemes have been considered. John (1969), for instance, suggested an alternative multivariate ratio estimator for population mean using an arbitrary design. John (1969) compared variance and computational procedures of his estimator to that of Olkin (1958) and observed that, upto a first order approximation, variances of the two estimators are the same. However, in terms of computational procedures, John (1969) noted that his estimator is easier to compute and use,

While extending the work of Olkin (1958), Ngesa et al. (2012) considered a stratified random sampling scheme with varying weights in each stratum and defined a multivariate ratio estimator for finite population total using two auxiliary variables. Using a simulated data, Ngesa et al. (2012) observed that the proposed estimator had a smaller bias compared to Olkin's (1958).

Instead of using the usual arithmetic mean, Malik and Singh (2012), on the other hand, used harmonic and geometric means in stratified random sampling with k strata to construct some improved multivariate ratio-type estimators. Using a real data, Malik and Singh (2012) observed that though the improved estimators had the same MSE's and harmonic mean as that of Olkin's (1958), the estimators were, however, less biased. Despite this improvement in the properties of multivariate ratio-type estimators, these improved estimators were constructed under the assumption of a positive correlation between the auxiliary characters and the study variable.

Kumar and Chhapparwal (2016) used a linear combination of two auxiliary variables to construct a generalized multivariate ratio and regression type estimator for population mean. Motivated by the multivariate chain ratio-type estimator expressed by Lu (2013), Kumar and Chhapparwal (2016) proposed an improved class of multivariate ratio-type estimator. Using empirical data, the constructed general class of multivariate ratio-type estimators performed better than previous multivariate estimators.

Singh et al. (2016) extended the work of Malik and Singh (2012) by considering known population proportion of two auxiliary attributes. Singh et al. (2016) made a similar observation that while the MSEs of Olkin (1958) estimator and estimators based on harmonic and geometric means are the same, the multivariate ratio-type estimator based on harmonic mean had the least bias.

These studies on multivariate ratio-type estimation of population mean and total suffer a common weakness that the constructed estimators are not only biased, but also fail to address the problem of non-response. Therefore, there is need to construct an unbiased multivariate ratio-type estimator and under non-response.

Apart from bias elimination, several studies have also suggested ways of minimizing sampling errors. For instance, Deming (1944) and Mahalanobis (1944) studied various types of errors and how to minimize them in sample surveys. From separate works, mathematical models for describing such errors were obtained and how the models could be used to minimize errors

were suggested. Studies that initially focused on identifying and describing sampling errors included, among others, Sukhatme and Seth (1952) and Hansen et al. (1953). Hartley and Ross (1954) not only studied sources of survey errors, but also proposed ratio estimation as a method of minimizing the errors. Other studies that have adopted ratio estimation methods to improve efficiency of constructed estimators include, among others, Okafor and Lee (2000), Singh and Kumar (2008), Shabbir and Gupta (2010), Shabbir and Saghir (2012), Singh and Sisodia (2014), Lone and Tailor (2015) and Zaman and Yilmaz (2017).

While looking at some particular cases, Rao (1991) evaluated and improved the precision of the Hartley-Ross (1954) unbiased ratio estimator and found that the proposed estimator is consistent and efficient relative to previous estimators such as those developed by Neyman (1934) and Cochran (1940). Hedayat and Sinha (1991) presented a convenient sampling strategy based on utilization of auxiliary information. Sarndal et al (1992) investigated the effect of the strength of correlation between Y and X in ratio estimation. The study revealed that precision of ratio estimation is improved when the linear regression of Y on X passes through the origin. Substantial surveys have been done on the use of auxiliary information to improve performance of estimators, including Upadhyaya and Singh (1999), Singh and Tailor (2003), Kadilar and Cingi (2006), Khoshnevisan et al. (2007) and Singh and Kumar (2011). Bahl and Tuteja (1991) also suggested an exponential form of the ratio estimator. Most of these previous studies on ratio estimation methods have, however, assumed homogeneous populations. Focus has not been on heterogeneous populations and this calls for the construction of ratio estimators under stratified sampling technique.

Stratified random sampling scheme is a two-step procedure. In the first step, a population consisting N units is divided into non-overlapping k homogeneous sub-populations each consisting of N_c units ($c = 1, 2, \dots, k$). These sub-populations are known as strata, while the population characteristic used in stratifying this heterogeneous population is called (*stratifying factor*). The second step involves obtaining a simple random sample without replacement from each stratum. Stratified random sampling has been preferred to simple random sampling since it yields estimates with high precision than in simple random sampling (Murthy, 1967; Cochran, 1977; Daroga & Chaudhary, 2002; Sampath & Ammani, 2010; Chaudhary & Kumar, 2015).

This advantage is due to the fact that stratification reduces heterogeneity in a population resulting to minimum within-stratum variations.

Despite this advantage of stratified random sampling, its high efficiency requires proper choice of stratum sample sizes. Proportional allocation and optimum allocation are two ways of sample size allocation that have extensively been discussed in literature (Cochran, 1977; Daroga & Chaudhary, 2002; Sampath & Ammani, 2010; Chaudhary & Kumar, 2015). In proportional allocation, sampling is done such that the sampling fraction n_c/N_c (for $c = 1, 2, \dots, k$) remains constant for all the strata. In optimum allocation, a stratum sample size n_c is chosen to minimize variance of an estimator for a fixed sample size or for a fixed sampling cost. Having obtained a suitable sample size, the overall finite population total is then estimated by obtaining the sum of all stratum population total estimates.

Ratio estimations under stratified random sampling scheme has resulted to the concepts of separate and combined ratio estimators for population total (Cochran, 1977; Daroga & Chaudhary, 2002). In separate ratio estimation method, individual ratio estimates of population total in each stratum is computed and cumulative totals of these stratum totals is then obtained. In combined ratio estimation method, the simple random sample estimates of Y and X using the stratum sample data is first obtained and the mean estimates are then used to obtain a combined ratio estimator for finite population total. In separate ratio estimation, the assumption is that there is variation in the stratum ratio estimates, while combined ratio estimation assumes that there is no significant variation in the stratum ratios.

Previous studies on combined ratio estimation have been done using two approaches. In the first approach, different ratio-type estimators have been combined to obtain a 'combined' ratio estimator, while the other approach has involved using a single combined ratio estimate for all the strata to obtain the overall estimator of a parameter in question (Shabbir & Saghir, 2012; Chaudhary & Kumar, 2015). Based on this distinction, former approach is not restricted to stratified random sampling scheme while the latter approach only applies in stratified random sampling.

In the recent past, much attention has been on the first approach where

ratio-type estimators have been coined by combining and improving existing ratio-type estimators. For instance, Singh and Vishwakarma (2005) used the usual combined ratio estimator of population mean suggested by Hansen, Hurwitz & Gurney, (1946) and the combined product estimator of population mean to construct a combined ratio-cum-product estimator of population mean. Also, from the usual combined ratio estimator, Diana (1993) constructed a general family of combined ratio estimators. Kayuncu and Kadilar (2010) extended the work of Diana (1993) by suggesting an improved family of combined ratio estimators of population mean.

Bahl and Tuteja (1991), on the other hand, used the usual exponential ratio estimator of population mean to construct an exponential ratio-type estimator, which was further studied and an exponential product-type estimator suggested. Singh et al. (2008) later considered the two estimators by Bahl and Tuteja (1991) and combined them to form a combined ratio-type estimator for population mean. Other studies that have been constructed by combining various exponential ratio estimator include Srivastava (1967), Kadilar and Cingi (2005), Kumar, Chaudhary and Kadilar (2009), Sharma et al. (2013) and Singh and Sharma (2014) among others.

Now, using the second definition of a combined ratio estimator, Wu (1985) considered variance estimations of the usual combined ratio estimator and the combined regression estimator and suggested a class of estimators of variance of the combined ratio estimator of population mean. In addition to Wu's study (1985), Saxena, Nigam and Shukla (1995) focused on estimating variance of the combined ratio estimator for population mean using balanced half samples. Though Saxena et al. (1995) studied properties of the suggested variance estimator, the study did not, however, suggest an unbiased combined ratio estimator.

In stratified random sampling, it is assumed that the knowledge of both strata sizes and possibility of drawing a sample from each stratum is available. However, this is not always the case since certain stratifying factors remain unknown until when sample units are selected. In such cases, a simple random sample is selected and then the sampled units are classified and treated as the usual stratified samples. This technique is referred to as post stratification (Cochran, 1977).

Other studies on the use of auxiliary variable include Housila, Singh and Kim (2010), who considered a case of two auxiliary variables in a two phase sampling procedure and observed that their proposed estimator was more efficient than the previously suggested estimators. Chaudhary, Malik, Singh and Singh (2013) used auxiliary information under non-response to construct a general family of estimators for estimating population mean in systematic sampling. Using mathematical problems in literature, Chaudhary et al. (2013) proposed estimator in the study had a better precision than previous estimators.

Surbramani and Kumarapandiyam (2013), in contrast, studied ratio estimation by considering a case when median of the auxiliary variable is known. In this study, it was found that the proposed estimator performed better than the previous ratio-type estimators. Sharma and Singh (2014) used two auxiliary variables under second order approximation to improve previous ratio estimators in simple random sampling without replacement.

Use of auxiliary information in surveys has also led to construction of regression estimators. Regression estimation is suitable in cases where the regression line of Y on X does not pass through the origin (Cochran, 1977). The main problem in regression estimation is therefore how to obtain optimal value(s) of regression coefficient such that regression estimator is not only unbiased, but also has a uniform minimum variance. In this method, values of the regression coefficients can be pre-assigned or optimal estimates can be obtained from sample data. In the latter case, the optimal value of the coefficient is computed from the covariance of X and Y divided by variance of X , an expression obtained using ordinary least square method.

Some studies that have considered regression estimation method include Montanari (1998), who studied properties of the generalized regression estimator of population mean. Montanari (1968) considered a multivariate regression estimator of population mean using q -dimensional auxiliary variable vector \mathbf{X} , having $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{qi})$ as the i -th population unit for $i = 1, 2, \dots, N$. In the suggested regression estimator, Montanari (1998) applied both Horvitz-Thompson estimation method and Yates-Grundy for-

mula to estimate variance of the regression coefficient vector using two approaches. In the first method, Montanari (1998) used both model-assisted approach and a linear regression super-population model and observed that the estimator has a minimum variance among a class of asymptotically design-based regression estimators.

In another approach, Montanari (1998) assumed that there is no linear combination of the entries $\widehat{\mathbf{X}}$ with a zero sampling variance (that is, $Var(\widehat{\mathbf{X}})$ is non-singular) and obtained a minimum variance vector estimator compared to previous regression estimators. Montanari (1998), however, recommended that there is need for further studies to analyze the stability of this optimal estimator.

Hamad, Haider and Hanif (2013) developed a regression ratio-type estimator with two auxiliary variable x and z for a two-phase sampling scheme and compared it with regression estimators by Robson (1957), Sukhatme (1962), Raj (1965), Mohanty (1967), Srivastava (1971), Mukerjee (1987) and Sammiuddin and Hanif (2007) and observed that the suggested estimator performs better than these previous estimators.

Most previous studies on regression estimation have considered estimation of finite population mean using complete data under SRSWOR. Though the problem of non-response has been of little focus in the recent past, Verma, Singh and Singh (2013) have, however, suggested a general class of regression estimator under non-response in the response variable using systematic sampling scheme. Using Hansen and Hurwitz (1946) approach to take care of the non-response, Verma et al. (2013) observed that the percent relative efficiency of their estimator decreases with increase in the response rate.

Olayiwola, Popoola and Bisira (2016) modified the usual regression estimator to include more than one auxiliary variables and suggested an estimator for double sampling in stratified random sampling. Olayiwola et al. (2013) observed that even though their estimator had high bias since it overestimates finite population mean, the estimator, however, had a less variance compared to existing estimators. Other studies that have regression estimation include Mukerjee, Rao and Vijayan (1987), Sahoo, Sahoo and Mohanty (1993) and Hanif, Shahbaz and Ahmad (2010) among others.

Despite these comprehensive studies on ratio estimation, focus on construction of unbiased ratio estimators is, however, still wanting. Also, there is need to explore how the constructed unbiased-type estimators can be used in cases of more than one auxiliary variable using either univariate or multivariate estimation procedures. Moreover, there is need to construct an unbiased ratio estimator of population mean or total and use it in regression estimation. For each of these loopholes, both separate and combined ratio estimations should be explored. Nevertheless, these gaps are filled in this study.

2.3 *Non-Response in Sample Surveys*

In sample surveys, non-response occurs when there is a failure to measure or to make observation on some units in the selected sample resulting to missing value(s) for that sample unit (Cochran, 1977; Daroga and Chaudhary, 2002; Ouma et al., 2010; Oyoo and Ouma, 2014). Usually, non-respondents have systematically different perception towards a survey than the responding lot. This scenario results to non-response bias. Response rates are thus used to measure the likelihood of non-response bias. Non-response can either be item-wise or unit-wise. In item non-response, a respondent refuses to answer one or more survey questions, while in unit non-response, a sampled respondent completely refuses to respond.

Item non-response could be due to sensitivity of survey question(a), accidental skip of question(s), failing to record response by the researcher and loss of data during data processing. Unit non-response could be due to failure or inability to locate a sampled unit, complete refusal to participate, loss of data and inability to participate, for example due to language barrier. Non-response can also be classified as ignorable and non-ignorable non-response. In ignorable non-response, if the cause of non-response is known, then adjusting the sampling strategy is used to take care of non-response bias. For non-ignorable non-response, even if the cause is known, adjusting the sampling strategy cannot eliminate non-response bias.

In sample surveys, non-response should not be overlooked since it reduces

representativeness of a sample, which further influences inferences made about a population from which the sample is obtained (Daroga and Chaudhary, 2002). As explained by Cochran (1977), non-response divides a study population into two 'strata', where the first 'stratum' represents population units for which observations would be made if the units are sampled, while the other stratum consist of population units for which there will be non-response.

To understand the effect of non-response on a sample estimate, let N_1 and N_2 be the respective number of population units in the responding and non-responding strata while n_1 and n_2 be the respective number of sample units in the responding and non-responding strata. Further, let the respective proportion of response and non-response groups be $W_1 = \frac{N_1}{N}$ and $W_2 = \frac{N_2}{N}$ with the corresponding sample and population mean pairs for *stratum 1* and *stratum 2* be \bar{y}_1, \bar{Y}_1 and \bar{y}_2, \bar{Y}_2 . If a simple random sample is drawn from the population and the sample mean is used to estimate population mean, then only data for the sample obtained from *stratum 1* will be used.

Addressing the problem of non-response involves determining whether the probability of non-response depends on the observed and/or unobserved data values. Using Y as the response variable and X as the auxiliary variable, another Bernoulli variable T can be defined to have the distribution function

$$\Pr(T=t) = \begin{cases} 1, & \text{if the sampled unit responds} \\ 0, & \text{otherwise} \end{cases}$$

Now, using Y , X and T , following three cases of missingness mechanisms can be defined:

Missing completely at random (MCAR): Here, the probability of non-response is independent of the values of Y and X . That is, the observed values of Y form a random sub-sample of the sample values of Y . Mathematically, MCAR is expressed as

$$\Pr(T = 0 \mid Y = y, X = x) = \Pr(T = 0).$$

Missing at random (MAR): In this case, the missingness depends on X but

not on Y so that MAR can be expressed as

$$Pr(T = 0 | Y = y, X = x) = Pr(T = 0 | X = x).$$

This implies that the observed values of Y form a random samples within sub-classes of X .

Missing not at random (MNAR): Here, the probability of non-response depends on both Y and X . MNAR is thus a case of non-ignorable non-response.

Based on the effect of non-response on the asymptotic properties of constructed estimators, appropriate correction measures should be taken to reduce non-response in sample surveys. Some methods that have been suggested in literature to take care of non-response survey sampling include data imputation, resampling, partial deletion, weight adjustment and sub-sampling.

Imputation refers to substitution of some value for missing data. Unit imputation involves data point substitution, whereas item imputation involves substitution of a component of data point (Broemeling, 2009). Hot-deck and cold-deck imputations are examples of repeated imputations (Broemeling, 2009). The former involves imputing a missing value from a randomly selected similar record, while the latter entails selecting donors from another set of data. Imputation procedures assume that there are minimal variations in observations of units from one population. This assumption makes imputation methods unsuitable in surveys involving large samples since homogeneity is not always automatic in large samples.

Daroga and Chaudhary (2002) defined resampling as conducting repeated sampling and observed that resampling is a suitable method of correcting non-response in large samples. Using empirical data, Raghunathan (2004) and Broemeling (2009) considered this result by Daroga and Chaudhary (2002) and observed that resampling offers more accurate method of correcting missing values than imputation. Previously, Lunneborg (2000) had only discussed bootstrapping and jackknifing as major techniques used in resampling. Using ordered sampling procedures, Lunneborg (2000) found that resampling has minimal bias and error variance compared to imputa-

tion procedures.

Broemeling (2009) explained that bootstrapping involves selecting multiple random samples with replacement and generating the distribution of estimates, from the selected samples, of the parameter to be estimated. In jackknifing, only subsets of available data are used (Raghunathan, 2004; Broemeling 2009). While comparing bootstrapping and jackknifing, Broemeling (2009) concludes that the latter is used in statistical inference for estimating bias and standard error of a statistic. That is, jackknifing involves systematic recomputation of the statistic estimate by leaving out one or more observations at a time from the sample set.

Raghunathan (2004) observed that bootstrapping provides a powerful way to estimate both variance and distribution of a point estimator, while jackknifing only provides variance of the point estimator. Thus, jackknife is a specialized method for estimating variances, while bootstrap assesses variance of a point estimator by first estimating its whole distribution. Bootstrapping is, nevertheless, preferred to jackknife because both variance and distribution of estimates is obtained at once, unlike jackknife which only yields variance of estimators. However, based on properties of the estimator and computation procedures, jackknife does not involve comprehensive computations and is easy to apply in empirical studies. Other individuals who had also studied bootstrapping and jackknifing using simulated data included Wu (1986) and Shao and Tu (1995).

Brick and Kalton (1996) defined partial deletion as a method of reducing available data so that a data set has no missing values. Partial deletion includes listwise deletion and pairwise deletion. Listwise deletion involves omitting cases with missing data, while, in pairwise deletion, each element in the inter-correlation matrix is estimated using all available data (Sarndal & Lundstrom, 2005).

In weight adjustment, sampled units are classified into some groups based on some auxiliary information. Then inside each group, each responding sampled unit is assigned some weight which is the inverse of the response rate of the corresponding category (Chang & Ferry, 2012). This assignment implies that higher weights are assigned to classes with low response rates and vice versa. Even though this method does not require filling in gaps in the data, its use is however pegged on the assumption that the probability of

non-response is the same for all units within a class.

Tsybakov (2009) investigated appropriateness of imputation, resampling, partial deletion and weight adjustment in correcting missing values. Using empirical results and ordered sampling procedures, Tsybakov (2009) observed that weight adjustment and resampling offer a suitable method of correcting non-response in sample surveys. For detailed illustration on the use of weight adjustment technique in correcting non-response under various sampling techniques see Chang and Ferry (2012) and Oyoo and Ouma (2014).

Sub-sampling method involves drawing a subset of the already sampled non-responding units. Sub-sampling is similar to multi-phase sampling where in the first phase, primary set of targeted respondents is obtained and in the second phase, secondary set from the primary non-responding units is obtained. This procedure begins by determining a sample size required to attain desired level of precision. In correcting non-response, sub-sampling method is based on the assumption that the sub-sample in the second phase will now have a complete data. Phase I sample units and the sub-sampled phase II units are then used to estimate population total. This method was suggested by Hansen and Hurwitz (1946).

Hansen-Hurwitz sub-sampling method has been widely used in literature while constructing various estimates of population parameters such as studies by Walsh (1970), Reddy (1973) and Srivastava (1967), Khoshnevisan et al. (2007), who constructed a general family of estimators for estimating population mean using known values of some population parameters, and Chaudhary and Kumar (2015).

Under non-response, Kumar (2012) utilized known population parameters to construct a general family of estimators of population mean. By varying the values of the constants in the estimator suggested by Kumar (2012), various estimators have been constructed. Saghir and Shabbir (2012), for instance, used Hansen-Hurwitz method to correct non-response while estimating finite population mean in stratified random sampling using auxiliary attribute. Similarly, Chaudhary et al. (2013) and Singh and Malik (2014) used this subsampling method to correct missing values while estimating finite population mean. Previously, Rao (1986) had studied the usual ratio

estimator for population mean under non-response using SRSWOR. Rao (1986) also obtained the bias and a large sample approximation to the MSE of the constructed estimator.

2.4 *Summary of Study Gaps*

This chapter has reviewed key areas in this study, which includes the use of auxiliary variable in ratio estimation, stratified random sampling and non-response. From the reviews, it can be noted that the problem of bias reduction and/or elimination in ratio estimation without tampering with its good properties still exists. Various solutions to this problem have been suggested with no conclusive agreement on how to address the problem. There is, therefore, need to construct an improved ratio-type estimator that perform better than previous ratio estimators. Another gap that has been observed is the little attention being given to multivariate ratio estimation since most of these previous studies, even those involving more than one auxiliary variable, have adopted univariate estimation procedures. There is also need to explore performance of the previously suggested unbiased ratio estimators using regression estimation approach. All these should be done while taking care of both separate and combined ratio estimations procedures and considering the problem of non-response in sample surveys.

3. METHODOLOGY

3.1 Introduction

This chapter outlines the various methods to be used in deriving various forms of unbiased ratio estimator for finite population total in stratified random sampling scheme under non-response. It further outlines the method to be used in construction of improved estimator.

3.2 Notation and Definition of Symbols

Consider a stratified population with N units consisting of k strata, where the c^{th} stratum has N_c units (for $c = 1, 2, \dots, k$) such that $\sum_{c=1}^k N_c = N$. From this population, we wish to select a simple random sample of size n without replacement. Using SRSWOR, a sample consisting of n_c units is selected from the c^{th} stratum such that $\sum_{c=1}^k n_c = n$. But under non-response, each stratum population units is divided into two disjoint groups of responding and non-responding units. Let subscript $j = 1$ denote responding group and $j = 2$ denote the non-responding group. Let Y_{cij} be the i^{th} population unit ($i = 1, 2, \dots, N_{cj}$) in group j ($j = 1, 2$) in stratum c ($c = 1, 2, \dots, k$). Using these definitions, the following notations are used:

$$Y_T = \sum_{c=1}^k \sum_{j=1}^2 \sum_{i=1}^{N_{cj}} Y_{cij} = \sum_{c=1}^k \sum_{j=1}^2 Y_{Tcj} = \sum_{c=1}^k Y_{Tc}: \text{overall population total and}$$

N_{cj} is the population size in c^{th} stratum in the j^{th} response group so that

$$N = \sum_{c=1}^k N_c = \sum_{c=1}^k \sum_{j=1}^2 N_{cj}.$$

$$\bar{Y}_{cj} = \frac{1}{N_{cj}} \sum_{i=1}^{N_{cj}} Y_{cij}: \text{population mean for the } j^{th} \text{ subgroup in stratum } c.$$

$\bar{Y}_c = \frac{1}{N_c} \sum_{j=1}^2 N_{cj} \bar{Y}_{cj}$: population mean for c^{th} stratum.

$\bar{Y} = \frac{1}{N} \sum_{c=1}^k N_c \bar{Y}_c$: overall population mean.

Similar notations are used for the auxiliary variable X .

$R_c = \frac{\bar{Y}_c}{\bar{X}_c}$: the usual population ratio in the c^{th} stratum .

$R_{cj} = \frac{\bar{Y}_{cj}}{\bar{X}_{cj}}$: the usual population ratio in c^{th} stratum in the j^{th} group.

$R_{cij} = \frac{Y_{cij}}{X_{cij}}$: i^{th} observation ratio in stratum c for j^{th} group so that

$$\bar{R}_{cj} = \frac{1}{N_{cj}} \sum_{i=1}^{N_{cj}} \frac{Y_{cij}}{X_{cij}}.$$

For variances, let,

$S_{yc}^2 = \frac{1}{N_c - 1} \sum_{i=1}^{N_c} (Y_{ci} - \bar{Y}_c)^2$: c^{th} stratum adjusted population variances for response variable, Y .

$S_{xc}^2 = \frac{1}{N_c - 1} \sum_{i=1}^{N_c} (X_{ci} - \bar{X}_c)^2$: c^{th} stratum adjusted population variances for auxiliary variable, X .

$S_{xycj} = \frac{1}{N_{cj} - 1} \sum_{i=1}^{N_{cj}} (X_{cij} - \bar{X}_{cj})(Y_{cij} - \bar{Y}_{cj})$: population co-variances between X and Y in stratum c for group j .

$S_{rxcj} = \frac{1}{N_{cj} - 1} \sum_{i=1}^{N_{cj}} (R_{cij} - \bar{R}_{cj})(X_{cij} - \bar{X}_{cj})$ population co-variances between R and X in stratum c for group j .

For stratum sample sizes, the corresponding lower cases n , n_c and n_{cj} have the definitions as above. However for the non-responding n_{c2} units, a sub-sample of m_c units is used to represent the units with incomplete data. For the totals and means, the corresponding lower cases for the sample shall be y_{tcj} , y_{tc} , y_t and \bar{y}_{cj} , \bar{y}_c , \bar{y} respectively. But under non-response, let

$y_{tc} = y_{tc1} + y_{tc2}$, where $y_{tc1} = n_{c1}\bar{y}_{c1}$ and $y_{tc2} = n_{c2}\bar{y}_{c2m}$ and \bar{y}_{c2m} is the mean obtained when m_c sub-sample units are used instead of n_{c2} units.

For any auxiliary variable X , similar notations and definitions shall apply, for both population characteristics and sample statistics. The corresponding sample ratios shall be $r_c = \frac{\bar{y}_c}{\bar{x}_c}$, $r_{cj} = \frac{\bar{y}_{cj}}{\bar{x}_{cj}}$, $r_{cij} = \frac{y_{cij}}{x_{cij}}$ so that $\bar{r}_{cj} = \frac{1}{n_{cj}} \sum_{i=1}^{n_{cj}} \frac{y_{cij}}{x_{cij}}$. The corresponding sample variances and co-variances shall be expressed using lower cases of the population variances and co-variances above.

3.3 Ratio Estimation of Population Total in Stratified Random Sampling

The use of auxiliary variable(s) to improve efficiency of estimators for various population parameters is not a new concept in sample surveys. This has been based on the assumption of known values of the auxiliary variable X . Using a known population mean for X as \bar{X} , the respective traditional ratio estimator (\bar{y}_R), product estimator (\bar{y}_P) and regression estimator (\bar{y}_{lr}) for population mean in SRSWOR are given as

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

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

$$\bar{y}_{lr} = \bar{y} + \mu(\bar{x} - \bar{X})$$

where μ is a constant determined such that $Var(\bar{y}_{lr})$ is minimum.

From these conventional estimators, several ratio-type estimators using various sampling schemes have been suggested. In stratified random sampling, we can either have separate or combined ratio estimators for both univariate and multivariate ratio estimations.

In separate ratio estimation method, we first obtain the estimates population totals in each stratum and then add these stratum totals (Cochran, 1977). In this case, we do not make assumption that the true ratio remains constant in all the strata. This estimation method, therefore, requires the knowledge of

the stratum means of the auxiliary variable \bar{X}_c .

By definition, the separate ratio estimator for finite population total \hat{Y}_{T_s} in stratified random sampling having k strata is

$$\hat{Y}_{T_s} = \sum_{c=1}^k N_c r_c \bar{X}_c = \sum_{c=1}^k \frac{\bar{y}_c}{\bar{x}_c} X_{Tc}, c = 1, 2, \dots, k \quad (3.1)$$

To a first order approximation, Cochran (1977) gives bias and variance of \hat{Y}_{T_s} as

$$\text{Bias}(\hat{Y}_{T_s}) = \sum_{c=1}^k \bar{Y}_c \left(\frac{N_c - n_c}{n_c} \right) \{C_{xc}^2 - \rho_c C_{xc} C_{yc}\} \quad (3.2)$$

and $\text{Var}(\hat{Y}_{T_s})$ is defined as

$$\text{Var}(\hat{Y}_{T_s}) = \sum_{c=1}^k \frac{N_c(N_c - n_c)}{n_c} (S_{yc}^2 + R_c^2 S_{xc}^2 - 2R_c \rho_c S_{xc} S_{yc}) \quad (3.3)$$

where C_{xc} is the coefficient of variation of X in stratum c and C_{yc} is the coefficient of variation of Y in stratum c

If the stratification is such that there are many strata (large k) and n_c is small, then $\text{Bias}(\hat{Y}_{T_s})$ will be significant relative to its standard error $\sigma(\hat{Y}_{T_s})$. This is true since in a particular stratum c , say, the relation

$$\frac{| \text{Bias}(Y_{Tsc}) |}{\sigma(\hat{Y}_{Tsc})} \leq C_{xc} \quad (3.4)$$

exists (Cochran, 1977).

In combined ratio estimation, the combined ratio is obtained using the ratio of the population totals. using SRSWOR, suppose \hat{Y}_{st} and \hat{X}_{st} denotes the respective population estimators for Y_T and X_T from a stratified sample as follows, then

$$\hat{Y}_{st} = \sum_{c=1}^k N_c \bar{y}_c \text{ and } \hat{X}_{st} = \sum_{c=1}^k N_c \bar{x}_c$$

so that the combined ratio estimate of finite population total is given by

$$\widehat{Y}_T = \frac{\widehat{Y}_{st}}{\widehat{X}_{st}} X_T = \frac{\bar{y}_{st}}{\bar{x}_{st}} X_T \quad (3.5)$$

where $\bar{y}_{st} = \frac{\widehat{Y}_{st}}{N}$, $\bar{x}_{st} = \frac{\widehat{X}_{st}}{N}$ are the respective estimates of population means of Y and X from the stratified sample.

Bias of \widehat{Y}_T to a first order approximation is given by

$$\text{Bias}(\widehat{Y}_T) = Y_T \sum_{c=1}^k \frac{n_c^2}{N_c^2} \frac{(N_c - n_c)}{n_c N_c} \left(\frac{S_{xc}^2}{\bar{X}^2} - \rho_c \frac{S_{xc}}{\bar{X}} \frac{S_{yc}}{\bar{Y}} \right) \quad (3.6)$$

which simplifies to

$$\text{Bias}(\widehat{Y}_T) = NR \left[\sum_{c=1}^k \frac{n_c^2}{N_c^2} \frac{(N_c - n_c)}{n_c N_c} S_{xc} \left(\frac{S_{xc}}{\bar{X}} - \rho_c \frac{S_{yc}}{\bar{Y}} \right) \right] \quad (\text{Cochran, 1977})$$

and $Var(\widehat{Y}_T)$ is given by

$$Var(\widehat{Y}_T) = \sum_{c=1}^k \frac{N_c^2 (N_c - n_c)}{n_c N_c} (S_{yc}^2 + R^2 S_{xc}^2 - 2R\rho_c S_{xc} S_{yc}) \quad (3.7)$$

(see Cochran, 1977)

Non-response in sample surveys occurs when there is a failure to measure or to make observation on some units in the selected sample and it divides study population into two disjoint 'strata' with the respective population and sample sizes N_1 , N_2 and n_1 , n_2 . Using SRSWOR, Hansen and Hurwitz (1946) suggested that from the n_2 non-respondents, a sub-sample of size $m = \frac{n_2}{h}$, $h \geq 1$ is drawn and it is assumed that the sub-sample has a complete data so that the sample mean pair for the auxiliary variable and the study variable can be denoted as $(\bar{x}_{2m}, \bar{y}_{2m})$. Using a single variable, Hansen and Hurwitz (1946) suggested an estimator for \bar{Y} as

$$\widehat{Y}_{HH} = w_1 \bar{y}_1 + w_2 \bar{y}_{2m} \quad (3.8)$$

Where $w_1 = \frac{n_1}{n}$, $w_2 = \frac{n_2}{n}$ and $n_1 + n_2 = n$

The estimator given in (3.8) is unbiased for $\bar{y} = w_1\bar{y}_1 + w_2\bar{y}_2$, which is further unbiased for

$$\bar{y}^* = \frac{1}{n} \sum_{i=1}^n y_i \quad (3.9)$$

Hansen and Hurwitz (1946) further expressed variance of \widehat{Y}_{HH} as

$$V(\widehat{Y}_{HH}) = \frac{N-n}{nN} S^2 + W_2 \frac{h-1}{n} S_2^2 \quad (3.10)$$

where $S^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{Y})^2$, $S_2^2 = \frac{1}{N_2-1} \sum_{i=1}^{N_2} (y_i - \bar{Y}_2)^2$, $W_2 = \frac{N_2}{N}$ and $h = \frac{n_2}{m}$.

By defining $s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2$ and $s_{2m}^2 = \frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y}_{2m})^2$ as the respective variances of the n_1 responding units and the m sub-sampled units, Hansen and Hurwitz (1946) expressed the unbiased estimator of \widehat{Y}_{HH} as

$$v(\widehat{Y}_{HH}) = \frac{N-n}{nN} \left[\frac{(n_1-1)s_1^2 + (n_2-h)s_{2m}^2}{n-1} + \frac{n_1(\bar{y}_1 - \widehat{Y}_{HH})^2}{n_2(\bar{y}_{2m} - \widehat{Y}_{HH})^2} \right] + \frac{w_2(N-1)(h-1)s_{2m}^2}{N(n-1)}$$

Since \widehat{Y}_{HH} is unbiased for \bar{Y} , then based on the Hansen-Hurwitz Method, the unbiased estimator for the finite population total is

$$\widehat{Y}_{HH} = N(w_1\bar{y}_1 + w_2\bar{y}_{2m}) \quad (3.11)$$

3.4 Regression Estimation

The regression estimator for population mean in SRSWOR is given by

$$\bar{y}_{lr} = \bar{y} + \mu(\bar{x} - \bar{X}) \quad (3.12)$$

where μ is a constant and is determined such that $V(\bar{y}_{lr})$ is minimum (Cochran, 1977).

From equation (3.12), it can be shown that

$$V(\bar{y}_{lr}) = V(\bar{y}) + \mu^2 V(\bar{x}) + 2\mu Cov(\bar{x}, \bar{y}) \quad (3.13)$$

so that using ordinary least square method, the estimator for μ is obtained as

$$\hat{\mu} = -\frac{Cov(\bar{x}, \bar{y})}{V(\bar{x})} = -\frac{S_{xy}}{S_x^2} \quad (3.14)$$

To estimate population total in SRSWOR, we multiply \bar{y}_{lr} by N and its variance by N^2 .

3.5 Multivariate Unbiased Ratio Estimation

A multivariate ratio estimator is constructed using a study variable Y and a p -dimensional auxiliary variable $\underline{\mathbf{X}}$ such that $\underline{\mathbf{X}} = (X_{1i}, X_{2i}, \dots, X_{pi})'$ where $\bar{\underline{\mathbf{X}}} \neq 0$.

In this case, let the subscript l , ($l = 1, 2, \dots, p$) denote the component of the random auxiliary vector $\underline{\mathbf{X}}$ so that $r_l = \frac{\bar{y}}{\bar{x}_l}$ becomes an unbiased estimator for $R_l = \frac{\bar{Y}}{\bar{X}_l}$, $l = 1, 2, \dots, p$.

Under SRSWOR, Olkin (1956) proposed a general multivariate form of the ratio estimator as

$$\hat{Y}_{MR} = W_1 \frac{\bar{y}}{\bar{x}_1} X_1 + W_2 \frac{\bar{y}}{\bar{x}_2} X_2 + \dots + W_p \frac{\bar{y}}{\bar{x}_p} X_p \quad (3.15)$$

where W_l 's (for $l = 1, 2, \dots, p$) are the weights that maximize the precision of \hat{Y}_{MR} subject to the linear condition that $\sum_{l=1}^p W_l = 1$

Now, equation (3.15) can also be expressed as

$$\hat{Y}_{MR} = W_1 r_1 X_1 + W_2 r_2 X_2 + \dots + W_p r_p X_p = \sum_{l=1}^p W_l r_l X_l = \sum_{l=1}^p W_l \hat{Y}_l \quad (3.16)$$

where \hat{Y}_l is the l^{th} component population total ratio estimate based on the l^{th} auxiliary variable component.

Clearly, from equations (3.15) and (3.16), the multivariate estimator suggested by Olkin (1956) is biased since it is based on a biased ratio estimator.

In stratified random sampling, Ngesa et al. (2012) defined a multivariate ratio estimator for finite population total as

$$\widehat{Y}_{MRE} = \sum_{c=1}^k \widehat{Y}_{MRc} \quad (3.17)$$

such that for the c^{th} stratum, we have

$$\widehat{Y}_{MRc} = W_{c1}\widehat{Y}_{Rc1} + W_{c2}\widehat{Y}_{Rc2} + \cdots + W_{cp}\widehat{Y}_{Rcp} \quad (3.18)$$

3.6 Weaknesses of Reviewed Estimators

Ratio estimators are known to perform better than estimators constructed under SRSWOR, especially when the regression line of Y on X passes through the origin. This property is further enhanced when stratified random sampling technique is used. However, ratio-type estimators constructed using the usual ratio often suffer a major weakness of biasness, a gap which this study addresses by constructing an unbiased ratio estimator for finite population total. Apart from construction of unbiased ratio estimator, ratio estimators constructed in this study differ greatly from the usual ratio estimator since the problem of non-response is addressed. In cases where the regression line of Y on X does not pass through the origin, ratio estimators constructed in this study are not only unbiased and address the problem of non-response, but also consider cases of lack of perfect linear relationship between X and Y .

Though the usual regression estimator is unbiased and is more efficient than the mean per unit estimator, it does not however give a solution to the problem of non-response, which is a gap addressed in this study. Even though Olkin (1956) and Ngesa et al. (2012) succeeded in constructing estimators when the auxiliary variable presents itself as a p -dimensional random vector, their estimators are still biased and do not address the problem of no-response. The ratio estimators in this study, therefore, deviates from the usual ratio estimators in literature by addressing the problems of bias, non-response and when Y is not perfectly correlated with X .

3.7 Construction of Improved Estimator

We have seen that the usual ratio estimator is biased in estimating finite population total. Also, under non-response, using the Hansen-Hurwitz sub-sampling method does not produce an unbiased ratio estimator in stratified random sampling. Our task is therefore to construct an unbiased ratio estimator for finite population total in stratified random sampling and use the Hansen-Hurwitz sub-sampling method to take care of the non-response. To eliminate the bias in the traditional ratio estimator, a mean ratio estimator

$$\bar{r}_1 = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i} = \frac{1}{n} \sum_{i=1}^n r_i$$

is considered and connect it to its bias by examining the population covariance of $\frac{y}{x}$ and x , where

$$\text{Cov} \left(\frac{y}{x}, x \right) = E \left(\frac{y}{x} \cdot x \right) - E \left(\frac{y}{x} \right) E(x).$$

Hartley and Ross (1954) considered this approach and obtained bias of \bar{r}_1 as

$$\text{Bias}(\bar{r}_1) = -\frac{1}{\bar{X}} \text{Cov} \left(\frac{y}{x}, x \right).$$

Now, using the expressions for \bar{r}_1 and $\text{Bias}(\bar{r}_1)$, Hartley and Ross (1954) obtained an unbiased ratio estimator as

$$r_u = \bar{r}_1 + \frac{n(N-1)}{(n-1)N\bar{X}} (\bar{y} - \bar{r}_1\bar{x}) \quad (3.19)$$

with a corresponding variance for large samples given as

$$S_{r_u}^2 = \frac{1}{n} (V(y) + R^2V(x) - 2RC(x, y)) \quad (3.20)$$

where $V(y)$ and $V(x)$ are respective population variances of Y and X while $C(x, y)$ is the population covariance of Y and X .

Now, from equation (3.19), the unbiased ratio estimator for population total in SRSWOR can be expressed as

$$\hat{Y}_T = \bar{r}_1 X_T + \frac{n(N-1)}{(n-1)} (\bar{y} - \bar{r}_1\bar{x}). \quad (3.21)$$

In this study, therefore, we shall construct an unbiased ratio estimator for finite population total by considering, in each stratum, the unbiased ratio form given in equation (3.19) and adopt the sub-sampling procedure suggested by Hansen and Hurwitz (1946). We shall repeat this procedure for all the strata by taking into consideration both separate and combined ratio estimation methods. The unbiased estimators obtained from this procedure shall be used to construct regression and multivariate unbiased ratio estimators.

3.8 Simulation Study

We use R to generate hypothetical data for simulation study. We consider a hypothetical population of 300 units. To obtain the stratum sizes, we randomly generate three values from uniform distribution such that the sum is 300. For one-auxiliary random data set, we generate normally distributed random vectors for the auxiliary variable X and for the response variable Y and fit a linear model of Y on X in each stratum using the linear regression model in R.

We use Krejcie-Morgan-Sample-Size-Table to get overall sample size as 170 and allocate the stratum sample sizes using proportional allocation technique. We assume a non-response rate of 20% in each stratum and partition the stratum population units accordingly.

We use random number generator to identify sample units in each response group in all the three strata. That is, sampling is done index-wise such that if i^{th} index is selected then the sample element will be the i^{th} pair (X_i, Y_i) . The paired sample data frames are then exported to excel for further computations.

For multi-auxiliary data simulations, we use the linear model

$$Y_i = \sum_{l=1}^p \beta_l X_{il} + e_i \quad (3.22)$$

where β'_j s are randomly generated from a uniform distribution while Y_i and

X_{il} are randomly generated from normal distribution with different parameters.

For the regression data simulation, we use the general linear regression model with an error term e , which follows a normal distribution. That is, we use the model

$$y = \omega x + e \quad (3.23)$$

where ω (constant term), x and y values are obtained in R as follows

$$\omega = rnorm(n, mean, sd)$$

$$x = runif(n, min., max.)$$

$$y = rnorm(n, mean, sd) \times x + rnorm(n, mean, sd)$$

and in each stratum, different set of values for the intercept term is used.

4. UNBIASED RATIO ESTIMATOR

4.1 Introduction

This chapter deals with construction of unbiased separate and combined ratio estimators for finite population total in stratified random sampling under non-response. Comparison of asymptotic properties with available ratio estimators is also done.

4.2 Unbiased Separate Ratio-Type Estimator

Under complete data, Hartley and Ross (1954), Cochran (1977) and Daroga and Chaudhary (2002) defined an unbiased ratio estimator for population mean in SRSWOR as

$$\widehat{Y} = \bar{r}\bar{X} + \frac{n(N-1)}{N(n-1)}(\bar{y} - \bar{r}\bar{x}) \quad (4.1)$$

where $\bar{r} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

In this section, we extend the estimator given in equation (4.1) to capture the case of stratification and non-response.

Under non-response, a ratio-type estimator for finite population total in stratified random sampling is suggested as

$$\widehat{Y}_T = \sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_{cj} X_{Tcj} + \frac{N_{cj}-1}{n_{cj}-1} (y_{tcj} - \bar{r}_{cj} x_{tcj}) \right] \quad (4.2)$$

where the notations and expressions have their usual meanings. The estimator given in equation (4.2) shall be denoted as \widehat{Y}_D .

Next, we show the derivation of \widehat{Y}_D .

For derivation of \widehat{Y}_D , consider a particular stratum c , say, partitioned into two disjoint groups. The first group consisting of responding and the other consisting of non-responding population units. By letting subscript $j = 1$ to denote responding group and $j = 2$ to denote the non-responding group and i to denote identity of a unit such that Y_{cij} is the i^{th} population unit ($i = 1, 2, \dots, N_{cj}$) in response group j ($j = 1, 2$) in stratum c ($c = 1, 2, \dots, k$) then for the two disjoint groups, the corresponding population totals are Y_{Tc1} and Y_{Tc2} so that

$$Y_{Tcj} = \sum_{i=1}^{N_{cj}} Y_{cij}, \text{ for } j = 1, 2$$

That is,

$$Y_T = \sum_{c=1}^k [\bar{R}_{c1} X_{Tc1} + \bar{R}_{c2} X_{Tc2}] \quad (4.3)$$

Under non-response, the usual ratio estimator for the finite population total in stratum c is given by

$$\widehat{Y}_{Tc} = \bar{r}_{c1} X_{Tc1} + \bar{r}_{c2} X_{Tc2}$$

where \bar{r}_{c1} and \bar{r}_{c2} are as previously defined.

But, $\text{Bias}(\widehat{Y}_{Tc}) = E(\widehat{Y}_{Tc}) - Y_{Tc}$,

which can be expanded as,

$$\text{Bias}(\widehat{Y}_{Tc}) = [X_{Tc1} E(\bar{r}_{c1}) - Y_{Tc1}] + [X_{Tc2} E(\bar{r}_{c2}) - Y_{Tc2}].$$

That is, $\text{Bias}(\widehat{Y}_{Tc}) = \text{Bias}(\widehat{Y}_{Tc1}) + \text{Bias}(\widehat{Y}_{Tc2})$

But in SRSWOR, we have,

$$\text{Cov}(\bar{x}_{c1}, \bar{y}_{c1}) = \frac{N_c - n_c}{n_c N_c} S_{xy c1}, \text{ Cov}(\bar{x}_{c2m}, \bar{y}_{c2m}) = \frac{n_{c2} - m_c}{m_c n_{c2}} S_{xy c2}$$

and

$$\text{Cov}(\bar{r}_{c1}, \bar{x}_{c1}) = \frac{N_c - n_c}{n_c N_c} S_{rx c1}, \text{ Cov}(\bar{r}_{c2}, \bar{x}_{c2m}) = \frac{n_{c2} - m_c}{m_c n_{c2}} S_{rx c2}$$

where, $S_{rxcj} = \frac{1}{N_{cj}-1} \sum_{i=1}^{N_{cj}} (R_{cij} - \bar{R}_{cj})(X_{cij} - \bar{X}_{cj})$, for $j = 1, 2$

which can further be expanded as follows,

$$\begin{aligned} S_{rxcj} &= \frac{1}{N_{cj}-1} \left[\sum_{i=1}^{N_{cj}} \frac{Y_{cij}}{X_{cij}} X_{cij} - N_{cj} \bar{R}_{cj} \bar{X}_{cj} \right] \\ &= \frac{1}{N_{cj}-1} \left[Y_{Tcj} - \bar{R}_{cj} X_{Tcj} \right] \end{aligned}$$

But \bar{r}_{cj} is unbiased for \bar{R}_{cj} such that $E(\bar{r}_{cj}) = \bar{R}_{cj}$ so that we can express S_{rxcj} as

$$\begin{aligned} S_{rxcj} &= \frac{1}{N_{cj}-1} \left[Y_{Tcj} - X_{Tcj} E(\bar{r}_{cj}) \right] \\ &= - \frac{1}{N_{cj}-1} \text{Bias}(\widehat{Y}_{Tcj}) \end{aligned}$$

an indication of inverse relationship between R and X .

Therefore, we can write covariance of \bar{r}_{c1} and \bar{x}_{c1} as

$$\text{Cov}(\bar{r}_{c1}, \bar{x}_{c1}) = - \frac{N_{c1} - n_{c1}}{n_{c1} N_{c1}} \frac{1}{N_{c1} - 1} \text{Bias}(\widehat{Y}_{Tc1}),$$

which implies that

$$\text{Bias}(\widehat{Y}_{Tc1}) = - \frac{n_{c1} N_{c1} (N_{c1} - 1)}{N_{c1} - n_{c1}} \text{Cov}(\bar{r}_{c1}, \bar{x}_{c1})$$

or equivalently

$$\text{Bias}(\widehat{Y}_{Tc1}) = - \frac{n_{c1} N_{c1} (N_{c1} - 1)}{N_{c1} - n_{c1}} \frac{N_{c1} - n_{c1}}{n_{c1} N_{c1}} S_{rxc1}.$$

which simplifies to,

$$\text{Bias}(\widehat{Y}_{Tc1}) = -(N_{c1} - 1) S_{rxc1} \quad (4.4)$$

Therefore, the estimator of the Bias of \widehat{Y}_{Tc1} is given by

$$\widehat{\text{Bias}}(\widehat{Y}_{Tc1}) = -(N_{c1} - 1) s_{rxc1},$$

where

$$\begin{aligned}
s_{rxc1} &= \frac{1}{n_{c1} - 1} \sum_{i=1}^{n_{c1}} (r_{ci1} - \bar{r}_{c1})(x_{ci1} - \bar{x}_{c1}) \\
&= \frac{1}{n_{c1} - 1} \left[\sum_{i=1}^{n_{c1}} r_{ci1}x_{ci1} - n_{c1}\bar{r}_{c1}\bar{x}_{c1} \right] \\
&= \frac{1}{n_{c1} - 1} \left[\sum_{i=1}^{n_{c1}} \frac{y_{ci1}}{x_{ci1}}x_{ci1} - n_{c1}\bar{r}_{c1}\bar{x}_{c1} \right] \\
&= \frac{1}{n_{c1} - 1} \left[\sum_{i=1}^{n_{c1}} y_{ci1} - n_{c1}\bar{r}_{c1}\bar{x}_{c1} \right] \\
&= \frac{1}{n_{c1} - 1} [y_{tc1} - \bar{r}_{c1}x_{tc1}]
\end{aligned}$$

We can thus express Bias in \widehat{Y}_{Tc1} as

$$\text{Bias}(\widehat{Y}_{Tc1}) = -\frac{(N_{c1} - 1)}{(n_{c1} - 1)} [y_{tc1} - \bar{r}_{c1}x_{tc1}]$$

Now using \widehat{Y}_{Tc1} as an estimator for Y_{Tc1} , we have

$$E[\widehat{Y}_{Tc1} - \text{Bias}(\widehat{Y}_{Tc1})] = Y_{Tc1}, \quad (4.5)$$

which implies that,

$$Y_{Tc1} = \widehat{Y}_{Tc1} + \frac{(N_{c1} - 1)}{(n_{c1} - 1)} [y_{tc1} - \bar{r}_{c1}x_{tc1}] \quad (4.6)$$

Using the same procedure for non-responding group, that is, for $j = 2$, we have

$$\widehat{\text{Bias}}(\widehat{Y}_{Tc2}) = -(n_{c2} - 1)s_{rxc2}$$

so that,

$$\text{Bias}(\widehat{Y}_{Tc2}) = -\frac{(n_{c2} - 1)}{(m_c - 1)} [y_{tc2} - \bar{r}_{c2}x_{tc2}]$$

Assuming proportional allocation of sample sizes in the responding groups such that $\frac{N_{c1}}{n_{c1}} \approx \frac{N_{c2}}{n_{c2}}$ and that there is a high response rate in the second sampling phase such that m_c , is so close to n_{c2} , then $\text{Bias}(\widehat{Y}_{Tc2})$ can be written as

$$\text{Bias}(\widehat{Y}_{Tc2}) = - \left(\frac{N_{c2}-1}{n_{c2}-1} \right) [y_{tc2} - \bar{r}_{c2}x_{tc2}]$$

Now, using the relation given in equation (4.5),

$$\widehat{Y}_{Tc2} = \left[\widehat{Y}_{Tc2} + \left(\frac{N_{c2}-1}{n_{c2}-1} \right) [y_{tc2} - \bar{r}_{c2}x_{tc2}] \right] \quad (4.7)$$

But $\widehat{Y}_{Tcj} = \bar{r}_{cj}X_{Tcj}$ so that using equation(4.6) and equation (4.7), the expression for \widehat{Y}_{Tc} becomes

$$\widehat{Y}_{Tc} = \sum_{j=1}^2 \left[\bar{r}_{cj}X_{Tcj} + \frac{(N_{cj}-1)}{(n_{cj}-1)} [y_{tcj} - \bar{r}_{cj}x_{tcj}] \right].$$

which can be summed over all strata to obtain Y_T as

$$\widehat{Y}_T = \sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_{cj}X_{Tcj} + \frac{N_{cj}-1}{n_{cj}-1} (y_{tcj} - \bar{r}_{cj}x_{tcj}) \right] = \widehat{Y}_D$$

Next, we study some asymptotic properties of \widehat{Y}_D . The following lemmas are used to show that the estimator \widehat{Y}_D is unbiased.

Lemma 4.1: The sample ratio mean for the j^{th} group in stratum c , \bar{r}_{cj} , is unbiased for the population ratio mean for the j^{th} group in stratum c , \bar{R}_{cj}

Proof. By definition, \bar{r}_{cj} is unbiased for \bar{R}_{cj} only if $E(\bar{r}_{cj}) = \bar{R}_{cj}$.

$$\text{Now, } E(\bar{r}_{cj}) = E \left(\frac{1}{n_{cj}} \sum_{i=1}^{n_{cj}} R_{cij} \right)$$

$$\text{That is, } E(\bar{r}_{cj}) = \frac{1}{n_{cj}} \sum_{i=1}^{n_{cj}} \frac{1}{N_{cj}} \sum_{i=1}^{N_{cj}} \frac{Y_{cij}}{X_{cij}},$$

which implies that,

$$E(\bar{r}_{cj}) = \frac{1}{n_{cj}} \sum_{i=1}^{n_{cj}} \bar{R}_{cj} = \frac{1}{n_{cj}} n_{cj} \bar{R}_{cj} = \bar{R}_{cj}$$

Hence the proof. □

Similarly, to show that the sample ratio mean in stratum c , \bar{r}_c , is unbiased for population ratio mean in stratum c , \bar{R}_c , the following lemma is used .

Lemma 4.2: The sample ratio mean in stratum c , \bar{r}_c , is unbiased for the population ratio mean in stratum c , \bar{R}_c .

Proof. By definition, \bar{r}_c is unbiased for \bar{R}_c only if $E(\bar{r}_c) = \bar{R}_c$.

$$\text{Now, } E(\bar{r}_c) = E\left(\frac{1}{n_c} \sum_{i=1}^{n_c} R_{ci}\right)$$

$$\text{That is, } E(\bar{r}_c) = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{Y_{ci}}{X_{ci}},$$

which implies that,

$$E(\bar{r}_c) = \frac{1}{n_c} \sum_{i=1}^{n_c} \bar{R}_c = \frac{1}{n_c} n_c \bar{R}_c = \bar{R}_c$$

Hence the proof. □

Theorem 4.1: The estimator \hat{Y}_D is an unbiased estimator for finite population total Y_T under a large sample assumption such that m_c , is very close to n_{c2}

This proof involves showing that $E(\hat{Y}_D) = Y_T$.

That is, showing that

$$E(\hat{Y}_D) = E\left(\sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_{cj} X_{Tcj} + \frac{N_{cj} - 1}{n_{cj} - 1} (y_{tcj} - \bar{r}_{cj} x_{tcj})\right]\right) = Y_T$$

Now, $E(\widehat{Y}_D)$ can be expanded as

$$E(\widehat{Y}_D) = \sum_{c=1}^k \left[E(\bar{r}_{c1}X_{Tc1} + \bar{r}_{c2}X_{Tc2}) + \frac{N_{c1} - 1}{n_{c1} - 1} E(y_{tc1} - \bar{r}_{c1}x_{tc1}) + \frac{N_{c2} - 1}{n_{c2} - 1} E(y_{tc2} - \bar{r}_{c2}x_{tc2}) \right] \quad (4.8)$$

From Lemma 4.1, $E(\bar{r}_{cj}X_{Tcj}) = \bar{R}_{cj}X_{Tcj}$ so that $E(\widehat{Y}_D)$ becomes

$$E(\widehat{Y}_D) = \sum_{c=1}^k \left[(\bar{R}_{c1}X_{Tc1} + \bar{R}_{c2}X_{Tc2}) + \frac{N_{c1} - 1}{n_{c1} - 1} E(y_{tc1} - \bar{r}_{c1}x_{tc1}) + \frac{N_{c2} - 1}{n_{c2} - 1} E(y_{tc2} - \bar{r}_{c2}x_{tc2}) \right] \quad (4.9)$$

Now,

$$\begin{aligned} y_{tc1} - \bar{r}_{c1}x_{tc1} &= n_{c1}\bar{y}_{c1} - n_{c1}\bar{r}_{c1}\bar{x}_{c1} \\ &= \sum_{i=1}^{n_{c1}} y_{ci1} - n_{c1}\bar{r}_{c1}\bar{x}_{c1} \\ &= \sum_{i=1}^{n_{c1}} \frac{y_{ci1}}{x_{ci1}} x_{ci1} - n_{c1}\bar{r}_{c1}\bar{x}_{c1} \\ &= \sum_{i=1}^{n_{c1}} (x_{ci1} - \bar{x}_{c1})(r_{ci1} - \bar{r}_{c1}) \end{aligned}$$

which reduces to,

$$y_{tc1} - \bar{r}_{c1}x_{tc1} = (n_{c1} - 1)s_{rxc1} \quad (4.10)$$

Similarly,

$$y_{tc2} - \bar{r}_{c2}x_{tc2} = n_{c2}\bar{y}_{c2} - n_{c2}\bar{r}_{c2}\bar{x}_{c2}$$

That is,

$$y_{tc2} - \bar{r}_{c2}x_{tc2} = n_{c2} \sum_{i=1}^{m_c} \frac{1}{m_c} y_{ci2} - m_c \bar{r}_{c2} \bar{x}_{c2} \quad (4.11)$$

Assuming that m_c is large and is close to n_{c2} such that $m_c \approx n_{c2}$, for all $c = 1, 2, \dots, k$, then $(y_{tc2} - \bar{r}_{c2}x_{tc2})$ in equation (4.11) can be simplified as

follows

$$\begin{aligned}
y_{tc2} - \bar{r}_{c2}x_{tc2} &= \sum_{i=1}^{n_{c2}} y_{ci2} - n_{c2}\bar{r}_{c2}\bar{x}_{c2} \\
&= \sum_{i=1}^{n_{c2}} \frac{y_{ci2}}{x_{ci2}} x_{ci2} - n_{c2}\bar{r}_{c2}\bar{x}_{c2} \\
&= \sum_{i=1}^{n_{c2}} (x_{ci2} - \bar{x}_{c2})(r_{ci2} - \bar{r}_{c2})
\end{aligned}$$

which can be expressed as,

$$y_{tc2} - \bar{r}_{c2}x_{tc2} = (n_{c2} - 1)s_{rxc2} \quad (4.12)$$

Therefore, substituting equation (4.10) and equation (4.12) in equation (4.9) and assuming that $m_c \approx n_{c2}$ gives,

$$\begin{aligned}
\hat{Y}_D &= \sum_{c=1}^k \left[(\bar{R}_{c1}X_{Tc1} + \bar{R}_{c2}X_{Tc2}) + \frac{(N_{c1} - 1)}{(n_{c1} - 1)}(n_{c1} - 1)(s_{rxc1}) \right. \\
&\quad \left. + \frac{(N_{c2} - 1)}{(n_{c2} - 1)}(n_{c2} - 1)(s_{rxc2}) \right]
\end{aligned}$$

Now,

$$\begin{aligned}
E(\hat{Y}_D) &= \sum_{c=1}^k \left[(\bar{R}_{c1}X_{Tc1} + \bar{R}_{c2}X_{Tc2}) + \frac{(N_{c1} - 1)}{(n_{c1} - 1)}E(n_{c1} - 1)(s_{rxc1}) \right. \\
&\quad \left. + \frac{(N_{c2} - 1)}{(n_{c2} - 1)}E(n_{c2} - 1)(s_{rxc2}) \right]
\end{aligned}$$

which simplifies to

$$E(\hat{Y}_D) = \sum_{c=1}^k [(\bar{R}_{c1}X_{Tc1} + \bar{R}_{c2}X_{Tc2}) + (N_{c1} - 1)E(s_{rxc1}) + (N_{c2} - 1)E(s_{rxc2})] \quad (4.13)$$

That is,

$$E(\hat{Y}_D) = \sum_{c=1}^k [(\bar{R}_{c1}X_{Tc1} + \bar{R}_{c2}X_{Tc2}) + (N_{c1} - 1)S_{rxc1} + (N_{c2} - 1)S_{rxc2}] \quad (4.14)$$

By definition,

$$S_{rxcj} = \frac{1}{N_{cj}-1} \sum_{i=1}^{N_{cj}} (X_{cij} - \bar{X}_{cj})(Y_{cij} - \bar{Y}_{cj}) = \frac{1}{N_{cj}-1} (Y_{Tcj} - N_{cj} \bar{R}_{cj} \bar{X}_{cj})$$

so that equation (4.14) becomes

$$E(\hat{Y}_D) = \sum_{c=1}^k [(\bar{R}_{c1} X_{Tc1} + \bar{R}_{c2} X_{Tc2}) + (Y_{Tc1} - N_{c1} \bar{R}_{c1} \bar{X}_{c1}) + (Y_{Tc2} - N_{c2} \bar{R}_{c2} \bar{X}_{c2})]$$

That is,

$$E(\hat{Y}_D) = \sum_{c=1}^k [Y_{Tc1} + Y_{Tc2}] = \sum_{c=1}^k Y_{Tc} = Y_T$$

Hence the proof.

For Mean Squared Error (MSE), the expression for MSE of \hat{Y}_D , by definition, is

$$MSE(\hat{Y}_D) = E[\hat{Y}_D - Y_T]^2$$

which can be expressed as

$$\begin{aligned} MSE(\hat{Y}_D) &= E[\hat{Y}_D + E(\hat{Y}_D) - E(\hat{Y}_D) - Y_T]^2 \\ &= E[\hat{Y}_D - E(\hat{Y}_D)]^2 + [E(\hat{Y}_D) - Y_T]^2 \\ &= Var(\hat{Y}_D) + [Bias(\hat{Y}_D)]^2 \end{aligned}$$

But under a large sample assumption and that $m_c \approx n_{c2}$, $Bias(\hat{Y}_D)$ asymptotically tends to 0, so that

$$MSE(\hat{Y}_D) = Var(\hat{Y}_D) \quad (4.15)$$

We now obtain the expression for $Var(\hat{Y}_D)$. We shall use the following theorem,

Theorem 4.2: Under large sample assumption, variance of \hat{Y}_D is given as

$$Var(\hat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{y_{cj}}^2 + \bar{R}_{cj}^2 S_{x_{cj}}^2 - 2\bar{R}_{cj} S_{xy_{cj}} + \frac{1}{n_{cj} - 1} [S_{rcj}^2 S_{x_{cj}}^2 + S_{rxcj}] \right] \quad (4.16)$$

Proof. In this proof, the results by Hartley and Ross (1954) and Goodman and Hartley (1958) on the properties of an unbiased ratio estimator are considered.

Hartley and Ross (1954) and Goodman and Hartley (1958) considered an unbiased ratio estimator

$$r_u = \bar{r}_1 + \frac{n(N-1)}{(n-1)N\bar{X}} (\bar{y} - \bar{r}_1\bar{x}) \text{ see equation (3.19)}$$

where $\bar{r}_1 = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}$ so that an unbiased ratio estimator for finite population total becomes

$$\hat{Y}_T = \bar{r}_1 X_T + \frac{n(N-1)}{(n-1)} (\bar{y} - \bar{r}_1\bar{x})$$

Goodman and Hartley (1958) obtained variance of r_u for a sample of any size as

$$Var(r_u) = \frac{1-f}{n\bar{X}^2} \left[S_y^2 + \bar{R}^2 S_x^2 - 2\bar{R}S_{xy} + \frac{1}{n-1} [S_r^2 S_x^2 + S_{rx}] \right], \quad (4.17)$$

where $R_i = \frac{Y_i}{X_i}$ is the population observation ratio, \bar{R} is the population mean of R_i and S_r^2 is the population variance of R_i , S_{rx} is the population covariance of R and X while $f = \frac{n}{N}$ is the sampling fraction.

To derive the expression for $Var(\hat{Y}_D)$, \hat{Y}_D is expressed as

$$\hat{Y}_D = \sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_{cj} X_{Tcj} + \frac{n_{cj}(N_{cj}-1)}{n_{cj}-1} (\bar{y}_{cj} - \bar{r}_{cj}\bar{x}_{cj}) \right] \quad (4.18)$$

Now, using $N\bar{X}r_u$ as an unbiased estimator for Y_T , then using r_u variance of the unbiased ratio estimator \hat{Y}_D for finite population total Y_T in SRSWOR is given as

$$Var(\hat{Y}_D) = \frac{N(N-n)}{n} \left[S_y^2 + \bar{R}^2 S_x^2 - 2\bar{R}S_{xy} + \frac{1}{n-1} [S_r^2 S_x^2 + S_{rx}] \right]$$

Therefore, in stratified random sampling and under non-response, we have $Var(\hat{Y}_D)$ given as

$$Var(\hat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj}-n_{cj})}{n_{cj}} \left[S_{y_{cj}}^2 + \bar{R}_{cj}^2 S_{x_{cj}}^2 - 2\bar{R}_{cj} S_{xy_{cj}} + \frac{1}{n_{cj}-1} [S_{r_{cj}}^2 S_{x_{cj}}^2 + S_{rx_{cj}}] \right] \quad (4.19)$$

which is the required proof. \square

Now, under a large sample assumption, the coefficient $\frac{1}{(n_{cj}-1)}$ becomes negligible so that equation (4.19) reduces to

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_{cj}^2 S_{xcj}^2 - 2\bar{R}_{cj} S_{xycj} \right], \quad (4.20)$$

Further, variance of the estimator in the c^{th} stratum in subgroup j can be expressed as

$$Var(\widehat{Y}_{Dcj}) = \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_{cj}^2 S_{xcj}^2 - 2\bar{R}_{cj} S_{xycj} + \frac{1}{n_{cj} - 1} [S_{rcj}^2 S_{xcj}^2 + S_{rxcj}] \right] \quad (4.21)$$

and the unbiased estimator of $Var(\widehat{Y}_{Dcj})$ becomes

$$Var(\widehat{Y}_{Dcj}) = \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[s_{ycj}^2 + \bar{r}_{cj}^2 s_{xcj}^2 - 2\bar{r}_{cj} s_{xycj} + \frac{1}{n_{cj} - 1} [s_{rcj}^2 s_{xcj}^2 + s_{rxcj}] \right] \quad (4.22)$$

where $s_{ycj}^2 = \frac{1}{n_{cj}-1} \sum_{i=1}^{n_{cj}} (y_{cij} - \bar{y}_{cj})^2$, $s_{xcj}^2 = \frac{1}{n_{cj}-1} \sum_{i=1}^{n_{cj}} (x_{cij} - \bar{x}_{cj})^2$ and $s_{rcj}^2 = \frac{1}{n_{cj}-1} \sum_{i=1}^{n_{cj}} (r_{cij} - \bar{r}_{cj})^2$ are the respective unbiased estimators for S_y^2 , S_x^2 and S_r^2 while $s_{xycj} = \frac{1}{n_{cj}-1} \sum_{i=1}^{n_{cj}} (x_{cij} - \bar{x}_{cj})(y_{cij} - \bar{y}_{cj})$ and $s_{rxcj} = \frac{1}{n_{cj}-1} \sum_{i=1}^{n_{cj}} (r_{cij} - \bar{r}_{cj})(x_{cij} - \bar{x}_{cj})$ are the respective estimators for S_{xy} and S_{rx} .

Now, $Var(\widehat{Y}_D)$ can further be expressed as

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_{cj}^2 S_{xcj}^2 - 2\bar{R}_{cj} \rho_{cj} S_{xcj} S_{ycj} \right] \quad (4.23)$$

where ρ_{cj} is the coefficient of correlation between X and Y while S_{xcj} and S_{ycj} are respective standard deviations of X and Y , all in stratum c , subgroup j .

From equation (4.23), we observe that $MSE(\widehat{Y}_D)$ decreases as the stratum sample sizes in the first sampling and the second sampling phases becomes large. Also, for a sufficiently large ρ_{cj} , $Var(\widehat{Y}_D)$ reduces significantly. That

is, $Var(\widehat{Y}_D)$ decreases when the regression line of Y on X is a straight line that passes through the origin, which is an assumption in ratio estimation.

Corollary 4.1: For large populations, and consequently large samples, $Var(\widehat{Y}_D)$ can be expressed as

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj} - n_{cj}}{n_{cj}} \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{R}_{cj} X_{cij})^2 \quad (4.24)$$

Proof. To prove Corollary 4.1, consider equation (4.20) and express S_{ycj}^2 , S_{xcj}^2 and S_{xycj} as follows

$$\begin{cases} S_{ycj}^2 = \frac{1}{N_{cj}-1} \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{Y}_{cj})^2 \\ S_{xcj}^2 = \frac{1}{N_{cj}-1} \sum_{i=1}^{N_{cj}} (X_{cij} - \bar{X}_{cj})^2 \\ S_{xycj} = \frac{1}{N_{cj}-1} \sum_{i=1}^{N_{cj}} (X_{cij} - \bar{X}_{cj}) (Y_{cij} - \bar{Y}_{cj}) \end{cases} \quad (4.25)$$

Using the expressions for S_{ycj}^2 , S_{xcj}^2 and S_{xycj} in equation set (4.25), $S_{ycj}^2 + \bar{R}_{cj}^2 S_{xcj}^2 - 2\bar{R}_{cj} S_{xycj}$ can be expanded as follows

$$\begin{aligned} S_{ycj}^2 + \bar{R}_{cj}^2 S_{xcj}^2 - 2\bar{R}_{cj} S_{xycj} &= \frac{1}{N_{cj}-1} \sum_{i=1}^{N_{cj}} \left\{ (Y_{cij} - \bar{Y}_{cj})^2 + \bar{R}_{cj}^2 (X_{cij} - \bar{X}_{cj})^2 \right. \\ &\quad \left. - 2\bar{R}_{cj} (X_{cij} - \bar{X}_{cj}) (Y_{cij} - \bar{Y}_{cj}) \right\} \\ &= \frac{1}{N_{cj}-1} \sum_{i=1}^{N_{cj}} [(Y_{cij} - \bar{Y}_{cj}) - \bar{R}_{cj} (X_{cij} - \bar{X}_{cj})]^2 \end{aligned}$$

So that the expression of $Var(Y_D)$ in equation (4.20) can be simplified as

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left\{ \frac{1}{N_{cj}-1} \sum_{i=1}^{N_{cj}} [(Y_{cij} - \bar{Y}_{cj}) - \bar{R}_{cj} (X_{cij} - \bar{X}_{cj})]^2 \right\} \quad (4.26)$$

That is,

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left\{ \frac{1}{N_{cj}-1} \sum_{i=1}^{N_{cj}} [Y_{cij} - \bar{Y}_{cj} - \bar{R}_{cj} X_{cij} + \bar{R}_{cj} \bar{X}_{cj}]^2 \right\},$$

which further reduces to,

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left\{ \frac{1}{N_{cj} - 1} \sum_{i=1}^{N_{cj}} [Y_{cij} - \bar{R}_{cj} X_{cij}]^2 \right\} \quad (4.27)$$

Now, for large populations, the ratio $\frac{N_{cj}}{N_{cj} - 1}$ asymptotically approaches 1 so that equation (4.27) reduces to

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj} - n_{cj}}{n_{cj}} \sum_{i=1}^{N_{cj}} [Y_{cij} - \bar{R}_{cj} X_{cij}]^2$$

Hence the proof. □

Corollary 4.1, implies that for small samples, we have

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left\{ \frac{1}{N_{cj} - 1} \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{R}_{cj} X_{cij})^2 + \frac{1}{n_{cj} - 1} [S_{rcj}^2 S_{xcj}^2 + S_{rxcj}] \right\}$$

In Corollary 4.1, the expression for $Var(\widehat{Y}_D)$ has been obtained under the assumption that, in the second sampling phase, m_c is so close to n_{c2} such that $\frac{n_{c2}}{m_c}$ tends to 1.

Corollary 4.2: Under proportional allocation,

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_j - n_j)}{n_j} \left\{ \frac{1}{N_{cj} - 1} \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{R}_{cj} X_{cij})^2 + \frac{N_j}{n_j N_{cj} - N_j} [S_{rcj}^2 S_{xcj}^2 + S_{rxcj}] \right\} \quad (4.28)$$

Proof. Under proportional allocation of sample sizes in various strata, we have

$$n_c = \frac{n N_c}{N}$$

However, under non-response, the allocated sample size in stratum c response group j is given by

$$n_{cj} = \frac{n_j N_{cj}}{N_j}$$

When substituted in (4.19) and simplified, $Var(\widehat{Y}_D)$ becomes

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_j - n_j)}{n_j} \left\{ \frac{1}{N_{cj} - 1} \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{R}_{cj} X_{cij})^2 + \frac{N_j}{n_j N_{cj} - N_j} [S_{rcj}^2 S_{xcj}^2 + S_{rxcj}] \right\}$$

Hence the proof. \square

For consistency, let $\{\widehat{Y}_D^*\}$ be the sequence of point estimators of finite population total. By definition, the sequence $\{\widehat{Y}_D^*\}$ is said to be weakly consistent for Y_T if \widehat{Y}_D^* converges in probability to Y_T as the sample size becomes large (Cochran, 1977).

Theorem 4.3: For a large population and consequently a large sample size, the unbiased ratio-type estimator \widehat{Y}_D is a consistent estimator of the finite population total Y_T .

Proof. Here, Chebyshev's inequality is used to prove that \widehat{Y}_D is a consistent estimator for Y_T .

From equation (4.19), $Var(\widehat{Y}_D)$ is expressed as

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left\{ \frac{1}{N_{cj} - 1} \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{R}_{cj} X_{cij})^2 + \frac{1}{n_{cj} - 1} [S_{rcj}^2 S_{xcj}^2 + S_{rxcj}] \right\}$$

For j^{th} response group, $Var(\widehat{Y}_{Dj})$ can be expressed as

$$Var(\widehat{Y}_{Dj}) = \sum_{c=1}^k \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left\{ \frac{1}{N_{cj} - 1} \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{R}_{cj} X_{cij})^2 + \frac{1}{n_{cj} - 1} [S_{rcj}^2 S_{xcj}^2 + S_{rxcj}] \right\} \quad (4.29)$$

Considering the expression for $Var(\widehat{Y}_{Dj})$ and assuming a large population and consequently a large sample, it can be shown that \widehat{Y}_{Dj} is a consistent

estimator for Y_{Tj} using Chebyshev's inequality. That is, as the sample size n_{cj} increases, the difference $N_{cj} - n_{cj}$ tends to zero such that

$$\lim_{n_{cj} \rightarrow N_{cj}} Pr\{|\widehat{Y}_{Dj} - Y_{Tj}| > \varepsilon\} = 0 \quad (4.30)$$

for every $\varepsilon > 0$

Now, using the Chebyshev's inequality, \widehat{Y}_{Dj} is a consistent estimator for Y_{Tj} if

$$Pr\{|\widehat{Y}_{Dj} - Y_{Tj}| > \varepsilon\} \leq \frac{Var(Y_{Dj})}{\varepsilon^2}$$

so that

$$\begin{aligned} Pr\{|\widehat{Y}_{Dj} - Y_{Tj}| > \varepsilon\} &\leq = \frac{Var(Y_{Dj})}{\varepsilon^2} \\ &= \frac{1}{\varepsilon^2} \sum_{c=1}^k \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left\{ \frac{1}{N_{cj} - 1} \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{R}_{cj} X_{cij})^2 \right. \\ &\quad \left. + \frac{1}{n_{cj} - 1} [S_{rcj}^2 S_{xcj}^2 + S_{rxcj}] \right\} \end{aligned} \quad (4.31)$$

Taking limits as $n_{cj} \rightarrow N_{cj}$, the right hand side of equation (4.31) tends to zero.

Hence, $\widehat{Y}_{Dj} \xrightarrow{p} Y_{Tj}$, which is the condition for consistency. Since \widehat{Y}_{Dj} is a consistent estimator for Y_{Tj} , it implies that \widehat{Y}_D is a consistent estimator for Y_T .

Hence the proof. □

Suppose in each stratum, the sample sizes for both phase I and phase II are large, then using the central limit theorem, the confidence interval of Y_T is given by

$$\widehat{Y}_D \pm Z_{\frac{\alpha}{2}} \sqrt{Var(\widehat{Y}_D)} \quad (4.32)$$

Where $Z_{\frac{\alpha}{2}}$ is the co-efficient value $(1 - \frac{\alpha}{2})$, which is obtained from the standard normal table at $(1 - \alpha)100\%$ confidence level.

Next is to show that \widehat{Y}_D is a best linear unbiased estimator for Y_T .

By definition, an estimator $\hat{\theta}$ is said to be a best linear unbiased estimator for an unknown parameter θ based on data X if it is unbiased such that $E[b'X] = \theta$, is a linear function of X such that the estimator can be written as $b'X$, where b' is a vector of constants, and has the smallest variance among all other unbiased linear estimators. (see Cochran, 1977).

Now, from the expression for \hat{Y}_D , \hat{Y}_D is a best linear unbiased estimator for Y_T if and only if \bar{r}_{cj} is a best linear unbiased estimator for \bar{R}_{cj} for $j = 1, 2$.

But from the optimality condition of the usual ratio estimator constructed under SRSWOR, the proposed estimator is the best linear unbiased estimator of finite population total when:

- (i) the linear relationship of y_{cij} on x_{cij} passes through the origin such that $y_{cij} = \beta_{cj}x_{cij} + \varepsilon_{cij}$, where ε'_{cij} s are independently and identically distributed with $E(\varepsilon_{cij}/x_{cij}) = 0$ and β_{cj} is the c^{th} - stratum slope parameter in partition j
- (ii) the line in (i) above is proportional to x_{cij} such that $Var(y_{cij}/x_{cij}) = E(\varepsilon_{cij}^2) = Dx_{cij}$, where D is a non-negative constant.

That is, \bar{r}_{cj} is a best linear unbiased estimator for \bar{R}_{cj} for $j = 1, 2$ and \hat{Y}_{Tcj} is the best linear unbiased estimator for Y_{Tcj} if, for a fixed x_{cj} , $E(y_{cij}) = \beta_{cj}x_{cj}$ and $Var(x_{cij}) \propto x_{cj}$, so that $Var(x_{cij}) = \pi_{cj}x_{cj}$, where π_{cj} is the proportionality constant in stratum c , response group j .

Let $\underline{y}_{cj} = (y_{c1j}, y_{c2j}, \dots, y_{cnj})'$ and $\underline{x}_{cj} = (x_{c1j}, x_{c2j}, \dots, x_{cnj})'$ denote two vectors of observations on y_{cij} 's and x_{cij} 's. Then, for a fixed \underline{x}_{cj} ,

- (i) $E(\underline{y}_{cj}) = \beta \underline{x}_{cj}$ and
- (ii) $Var(\underline{y}_{cj}) = \Pi = \pi \text{diag}(x_{c1j}, x_{c2j}, \dots, x_{cnj})$, where $\text{diag}(x_{c1j}, x_{c2j}, \dots, x_{cnj})$ is the diagonal matrix with $x_{c1j}, x_{c2j}, \dots, x_{cnj}$ as the diagonal elements.

Therefore, the best linear unbiased estimator of β_{cj} is obtained by minimizing the function

$$S_{cj}^2 = (\underline{y}_{cj} - \beta_{cj}\underline{x}_{cj})'\Pi^{-1}(\underline{y}_{cj} - \beta_{cj}\underline{x}_{cj})$$

which can be expressed as

$$S_{cj}^2 = \sum_{i=1}^{n_{cj}} \frac{(y_{cij} - \beta_{cj}x_{cij})^2}{\pi_{cj}x_{cij}} \quad (4.33)$$

Differentiating equation (4.33) with respect to β_{cj} and equating to 0 gives

$$\sum_{i=1}^{n_{cj}} (y_{cij} - \hat{\beta}_{cj}x_{cij}) = 0 \quad (4.34)$$

which implies that $\hat{\beta}_{cj} = \frac{1}{n_{cj}} \sum_{i=1}^{n_{cj}} \frac{y_{cij}}{x_{cij}} = \bar{r}_{cj}$. Hence, \bar{r}_{cj} is the linear unbiased estimator of \bar{R}_{cj} .

Since \bar{r}_{cj} is a best linear unbiased estimator for \bar{R}_{cj} for $j = 1, 2$, it implies that \hat{Y}_{Dcj} is the BLUE of Y_{Tcj} and consequently, \hat{Y}_D is the BLUE of Y_T .

We now compare variance of \hat{Y}_D under proportional allocation and under SRSWOR.

Theorem 4.4: Under a large sample assumption, the estimator \hat{Y}_D is more efficient under proportional allocation of sample sizes than when SRSWOR sampling scheme is used.

Proof. Let the variance of \hat{Y}_D under SRSWOR be denoted as $Var(\hat{Y}_D)_S$ and variance of \hat{Y}_D under proportional allocation be denoted as $Var(\hat{Y}_D)_{prop}$. Therefore, this proof involves comparing $Var(\hat{Y}_D)_{prop}$ and $Var(\hat{Y}_D)_S$, where

$$Var(\hat{Y}_D)_S = \sum_{j=1}^2 \frac{N_j(N_j - n_j)}{n_j} \left\{ \frac{1}{N_j - 1} \sum_{i=1}^{N_j} (Y_{ij} - \bar{R}_j X_{ij})^2 + \frac{1}{n_j - 1} [S_{rj}^2 S_{xj}^2 + S_{rxj}] \right\}$$

and

$$Var(\widehat{Y}_D)_{prop} = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_j - n_j)}{n_j} \left\{ \frac{1}{N_{cj} - 1} \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{R}_{cj} X_{cij})^2 + \frac{N_j}{n_j N_{cj} - N_j} [S_{rcj}^2 S_{xcj}^2 + S_{rcxj}] \right\}$$

For j^{th} response group, $Var(\widehat{Y}_{Dj})_{prop}$ and $Var(\widehat{Y}_{Dj})_S$ can be expressed as,

$$\begin{cases} Var(\widehat{Y}_{Dj})_S = \frac{N_j(N_j - n_j)}{n_j} \left\{ S_{yj}^2 + \frac{1}{n_j - 1} [S_{rj}^2 S_{xj}^2 + S_{rxj}] \right\} \\ Var(\widehat{Y}_{Dj})_{prop} = \frac{(N_j - n_j)}{n_j} \sum_{c=1}^k N_{cj} \left\{ S_{cyj}^2 + \frac{N_j}{n_j N_{cj} - N_j} [S_{crj}^2 S_{cxj}^2 + S_{crxj}] \right\} \end{cases} \quad (4.35)$$

Since the constant coefficient $\frac{(N_j - n_j)}{n_j}$ is common in both $Var(\widehat{Y}_{Dj})_{prop}$ and $Var(\widehat{Y}_{Dj})_S$, equation (4.35) thus reduces to

$$\begin{cases} N_j \left\{ S_{yj}^2 + \frac{1}{n_j - 1} [S_{rj}^2 S_{xj}^2 + S_{rxj}] \right\} \\ \left\{ \sum_{c=1}^k N_{cj} \left\{ S_{cyj}^2 + \frac{N_j}{n_j N_{cj} - N_j} [S_{crj}^2 S_{cxj}^2 + S_{crxj}] \right\} \right\} \end{cases}$$

For large samples such that the ratio $\frac{1}{n_{cj} - 1}$ is very close to $\frac{1}{n_{cj}}$ and the ratio $\frac{1}{n_j - 1}$ is very close to $\frac{1}{n_j}$, expressions for $Var(\widehat{Y}_{Dj})_{prop}$ and $Var(\widehat{Y}_{Dj})_S$ can be simplified and partitioned as

$$\begin{cases} Var(\widehat{Y}_{Dj})_S = \underbrace{N_j S_{yj}^2}_A + \underbrace{\frac{N_j}{n_j} [S_{rj}^2 S_{xj}^2 + S_{rxj}]}_B \\ Var(\widehat{Y}_{Dj})_{prop} = \underbrace{\sum_{c=1}^k N_{cj} S_{cyj}^2}_{A'} + \underbrace{N_j \sum_{c=1}^k \frac{1}{n_{cj}} [S_{crj}^2 S_{cxj}^2 + S_{crxj}]}_{B'} \end{cases} \quad (4.36)$$

The task is therefore to compare the corresponding partitions.

Now, from equation set (4.36), we have $S_{yj}^2 = \frac{1}{N_j - 1} \sum_{c=1}^k \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{Y}_j)^2$, which can be expanded and simplified as

$$\begin{aligned}
S_{yj}^2 &= \frac{1}{N_j - 1} \sum_{c=1}^k \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{Y}_{cj} + \bar{Y}_{cj} - \bar{Y}_j)^2 \\
&= \frac{1}{N_j - 1} \sum_{c=1}^k \sum_{i=1}^{N_{cj}} [(Y_{cij} - \bar{Y}_{cj}) + (\bar{Y}_{cj} - \bar{Y}_j)]^2 \\
&= \frac{1}{N_j - 1} \sum_{c=1}^k \sum_{i=1}^{N_{cj}} \left\{ (Y_{cij} - \bar{Y}_{cj})^2 + (\bar{Y}_{cj} - \bar{Y}_j)^2 + 2(Y_{cij} - \bar{Y}_{cj})(\bar{Y}_{cj} - \bar{Y}_j) \right\} \\
&= \frac{1}{N_j - 1} \sum_{c=1}^k \left\{ (N_{cj} - 1) S_{cyj}^2 + N_{cj} (\bar{Y}_{cj} - \bar{Y}_j)^2 \right\}
\end{aligned}$$

But for large population sizes, the ratio $\frac{1}{N_{cj}-1}$ is so close to $\frac{1}{N_{cj}}$ and the ratio $\frac{1}{N_j-1}$ tends to $\frac{1}{N_j}$ so that S_{yj}^2 can be expressed as

$$S_{yj}^2 = \frac{1}{N_j} \left\{ \sum_{c=1}^k N_{cj} S_{cyj}^2 + \sum_{c=1}^k N_{cj} (\bar{Y}_{cj} - \bar{Y}_j)^2 \right\} \quad (4.37)$$

Similarly,

$$\begin{aligned}
S_{rj}^2 &= \frac{1}{N_j - 1} \sum_{c=1}^k \sum_{i=1}^{N_{cj}} (R_{cij} - \bar{R}_{cj} + \bar{R}_{cj} - \bar{R}_j)^2 \\
&= \frac{1}{N_j - 1} \sum_{c=1}^k \left\{ (N_{cj} - 1) S_{crj}^2 + N_{cj} (\bar{R}_{cj} - \bar{R}_j)^2 \right\}
\end{aligned}$$

But for large population sizes sizes, the ratio $\frac{1}{N_{cj}-1}$ is so close to $\frac{1}{N_{cj}}$ and the ratio $\frac{1}{N_j-1}$ tends to $\frac{1}{N_j}$ so that S_{rj}^2 can expressed as

$$S_{rj}^2 = \frac{1}{N_j} \left\{ \sum_{c=1}^k N_{cj} S_{crj}^2 + \sum_{c=1}^k N_{cj} (\bar{R}_{cj} - \bar{R}_j)^2 \right\} \quad (4.38)$$

and

$$\begin{aligned}
S_{xj}^2 &= \frac{1}{N_j - 1} \sum_{c=1}^k \sum_{i=1}^{N_{cj}} (X_{cij} - \bar{X}_{cj} + \bar{X}_{cj} - \bar{X}_j)^2 \\
&= \frac{1}{N_j - 1} \sum_{c=1}^k \left\{ (N_{cj} - 1) S_{cxj}^2 + N_{cj} (\bar{X}_{cj} - \bar{X}_j)^2 \right\}
\end{aligned}$$

Similarly, for large population sizes such that the ratios $\frac{1}{N_{cj}-1}$ is so close to $\frac{1}{N_{cj}}$ and $\frac{1}{N_j-1}$ tends to $\frac{1}{N_j}$, then S_{xj}^2 becomes

$$S_{xj}^2 = \frac{1}{N_j} \left\{ \sum_{c=1}^k N_{cj} S_{cxj}^2 + \sum_{c=1}^k N_{cj} (\bar{X}_{cj} - \bar{X}_j)^2 \right\} \quad (4.39)$$

Now substituting equation (4.37) in equation (4.36) gives

$$A = \left\{ \sum_{c=1}^k N_{cj} S_{cyj}^2 + \sum_{c=1}^k N_{cj} (\bar{Y}_{cj} - \bar{Y}_j)^2 \right\},$$

so that $\Delta = A - A'$ becomes

$$\begin{aligned} \Delta &= \sum_{c=1}^k \left\{ N_{cj} S_{cyj}^2 + N_{cj} (\bar{Y}_{cj} - \bar{Y}_j)^2 - N_{cj} S_{cyj}^2 \right\} \\ &= \sum_{c=1}^k N_{cj} (\bar{Y}_{cj} - \bar{Y}_j)^2, \text{ which is a positive constant} \end{aligned}$$

Clearly, $\Delta > 0$ for all values of $c = 1, 2, \dots, k$ and $j = 1, 2$.

To compare B and B' , the difference can be expressed as

$$\begin{aligned} \Delta^* = B - B' &= \frac{1}{n_j} [S_{rj}^2 S_{xj}^2 + S_{rxj}] - \sum_{c=1}^k \frac{1}{n_{cj}} [S_{crj}^2 S_{cxj}^2 + S_{crxj}] \\ &= \frac{1}{n_j} [S_{rj}^2 S_{xj}^2 + S_{rxj}] - \sum_{c=1}^k \frac{1}{n_{cj}} [S_{crj}^2 S_{cxj}^2 + S_{crj} S_{cxj}] \end{aligned}$$

But in sample surveys, stratification reduces variance of estimators for population parameters such that the constant $\frac{1}{n_j} [S_{rj}^2 S_{xj}^2 + S_{rxj}]$ is far much greater than $\sum_{c=1}^k \frac{1}{n_{cj}} [S_{crj}^2 S_{cxj}^2 + S_{crj} S_{cxj}]$ so that, $\Delta^* > 0$.

Therefore, since both Δ and Δ^* are greater than zero, it implies that the difference $Var(\hat{Y}_{Dj})_S - Var(\hat{Y}_{Dj})_{prop}$ gives a positive constant, which further implies that $Var(\hat{Y}_D)_S - Var(\hat{Y}_D)_{prop} > 0$. It can thus be concluded that the estimator Y_D is more efficient under proportional allocation than under

SRSWOR.

Hence the proof. □

4.3 Combined Ratio Form of \widehat{Y}_D

Use of combined ratio estimation in stratified random sampling date back to 1946, where instead of using ratio estimates in each stratum to obtain the overall estimate of the population total, a single combined ratio estimator is derived for all the strata and used in estimation (Hansen, Hurwitz and Gurney, 1946; Cochran, 1977). Use of combined ratio estimation method is due to the fact that if separate ratio estimation method were to be used in the estimation problem, the accumulated bias, when aggregated over all strata, can be quite significant and consequently reducing efficiency of the estimator. One possible solution to reducing this accumulated bias is the use of combined ratio estimation method. That is, combined ratio estimator is much less prone to the risk of bias compared to separate ratio estimator.

In combined ratio estimation, the standard estimators for Y_T and X_T are first obtained as $\widehat{Y}_{st} = \sum_{c=1}^k N_c \bar{y}_c$, and $\widehat{X}_{st} = \sum_{c=1}^k N_c \bar{x}_c$ respectively. Using these estimators, the estimator for finite population total is then obtained as

$$\widehat{Y}_T = \frac{\widehat{Y}_{st}}{\widehat{X}_{st}} X_T = \frac{\bar{y}_{st}}{\bar{x}_{st}} X_T \quad (4.40)$$

where $\bar{y}_{st} = \frac{\widehat{Y}_{st}}{N}$, $\bar{x}_{st} = \frac{\widehat{X}_{st}}{N}$ are the respective estimators for population means of Y and X from the stratified sample and the estimator of the combined ratio $R = \frac{\bar{Y}}{\bar{X}}$ is defined as $\widehat{R} = \frac{\bar{y}_{st}}{\bar{x}_{st}}$.

Cochran (1977) gives the expressions of $\text{Bias}(\widehat{Y}_T)$ and $\text{Var}(\widehat{Y}_T)$ under combined ratio estimation, to a first order approximation, as

$$\text{Bias}(\widehat{Y}_T) = Y_T \sum_{c=1}^k \frac{n_c^2}{N_c^2} \frac{(N_c - n_c)}{n_c N_c} \left(\frac{S_{xc}^2}{\bar{X}^2} - \rho_c \frac{S_{xc}}{\bar{X}} \frac{S_{yc}}{\bar{Y}} \right) \quad (4.41)$$

and

$$Var(\widehat{Y}_T) = \sum_{c=1}^k \frac{N_c^2(N_c - n_c)}{n_c N_c} (S_{yc}^2 + R^2 S_{xc}^2 - 2R\rho_c S_{xc} S_{yc}) \quad (4.42)$$

where, $S_{yc}^2 = \frac{1}{N_c-1} \sum_{i=1}^{N_c} (Y_{ci} - \bar{Y}_c)^2$ and $S_{xc}^2 = \frac{1}{N_c-1} \sum_{i=1}^{N_c} (X_{ci} - \bar{X}_c)^2$ are the stratum adjusted population variances for the response variable Y and auxiliary variable X and ρ_c is the population correlation coefficient between X and Y in stratum c

In this study, R^* is used to denote the population combined ratio and is defined as

$$R^* = \frac{\sum_{c=1}^k Y_{Tc}}{\sum_{c=1}^k X_{Tc}} = \frac{\sum_{c=1}^k N_c \bar{Y}_c}{\sum_{c=1}^k N_c \bar{X}_c} \quad (4.43)$$

with a corresponding sample combined ratio expressed as

$$r^* = \frac{\sum_{c=1}^k n_c \bar{y}_c}{\sum_{c=1}^k n_c \bar{x}_c} \quad (4.44)$$

However, under non-response, the population combined ratio R^* is partitioned as

$$R_1^* = \frac{\sum_{c=1}^k N_{c1} \bar{Y}_{c1}}{\sum_{c=1}^k N_{c1} \bar{X}_{c1}} \quad \text{and} \quad R_2^* = \frac{\sum_{c=1}^k N_{c2} \bar{Y}_{c2}}{\sum_{c=1}^k N_{c2} \bar{X}_{c2}}$$

and the sample combined ratio r^* as

$$r_1^* = \frac{\sum_{c=1}^k n_{c1} \bar{y}_{c1}}{\sum_{c=1}^k n_{c1} \bar{x}_{c1}} \quad \text{and} \quad r_2^* = \frac{\sum_{c=1}^k m_{c2} \bar{y}_{c2m}}{\sum_{c=1}^k m_{c2} \bar{x}_{c2m}}$$

Assuming large samples such that the fraction $\frac{1}{m_c}$ is very close to $\frac{1}{n_{c2}}$ so that

$$\bar{y}_{c2m} \text{ tends to } \bar{y}_{c2}, \text{ then } r_2^* = \frac{\sum_{c=1}^k n_{c2} \bar{y}_{c2}}{\sum_{c=1}^k n_{c2} \bar{x}_{c2}}$$

Now, using $r_{ij}^* = \frac{y_{ij}}{x_{ij}}$, $i = 1, 2, \dots, n_j$, $j = 1, 2$ and $R_{ij}^* = \frac{Y_{ij}}{X_{ij}}$, $i = 1, 2, \dots, N_j$, $j = 1, 2$ as the sample and population observation ratios respectively, then the sample mean observation ratio is defined as

$$\bar{r}^* = \frac{1}{n} \sum_{j=1}^2 \sum_{i=1}^{n_j} \frac{y_{ij}}{x_{ij}} \quad (4.45)$$

and considered as an estimator for

$$\bar{R}^* = \frac{1}{N} \sum_{j=1}^2 \sum_{i=1}^{N_j} \frac{Y_{ij}}{X_{ij}} \quad (4.46)$$

Similarly, $\bar{r}_j^* = \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{y_{ij}}{x_{ij}}$ is an estimator for $\bar{R}_j^* = \frac{1}{N_j} \sum_{i=1}^{N_j} \frac{Y_{ij}}{X_{ij}}$.

The respective estimators for $S_{xyj} = \frac{1}{N_j-1} \sum_{i=1}^{N_j} (X_{ij} - \bar{X}_j)(Y_{ij} - \bar{Y}_j)$ and

$$S_{rxj} = \frac{1}{N_j-1} \sum_{i=1}^{N_j} (R_{ij} - \bar{R}_j^*)(X_{ij} - \bar{X}_j) \text{ are } s_{xyj} = \frac{1}{n_j-1} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j)$$

$$\text{and } s_{rxj} = \frac{1}{n_j-1} \sum_{i=1}^{n_j} (r_{ij} - \bar{r}_j^*)(x_{ij} - \bar{x}_j).$$

Using this notations, a combined ratio estimator of finite population total under non-response is thus suggested as,

$$\hat{Y}_{DC} = \sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_j^* X_{Tcj} + \frac{N_{cj} - 1}{n_{cj} - 1} \{y_{tcj} - \bar{r}_j^* x_{tcj}\} \right] \quad (4.47)$$

Next is to show how the combined ratio estimator \hat{Y}_{DC} is constructed. In SRSWOR, the usual ratio estimator $r = \frac{\bar{y}}{\bar{x}}$ is used as the estimator for the population ratio $R = \frac{\bar{Y}}{\bar{X}}$. Similarly, in stratified random sampling, $r_c = \frac{\bar{y}_c}{\bar{x}_c}$ is used to estimate the stratum population ratio $R_c = \frac{\bar{Y}_c}{\bar{X}_c}$ so that $\hat{Y}_c = r_c \bar{X}_c$ or equivalently, $\hat{Y}_{Tc} = N_c r_c \bar{X}_c = \sum_{c=1}^k r_c X_{Tc}$, where $c = 1, 2, \dots, k$.

However, in combined ratio estimation, the sample combined ratio

$$r^* = \frac{\sum_{c=1}^k n_c \bar{y}_c}{\sum_{c=1}^k n_c \bar{x}_c} \quad (4.48)$$

is used as an estimator for

$$R^* = \frac{\sum_{c=1}^k Y_{Tc}}{\sum_{c=1}^k X_{Tc}} = \frac{\sum_{c=1}^k N_c \bar{Y}_c}{\sum_{c=1}^k N_c \bar{X}_c} \quad (4.49)$$

Now, using $r_{ij}^* = \frac{y_{ij}}{x_{ij}}$ ($i = 1, 2, \dots, n_j, j = 1, 2$) and \bar{r}^* , estimator for Y_T can expressed as

$$\hat{Y}_T = \sum_{c=1}^k \sum_{j=1}^2 \hat{Y}_{Tcj} = \sum_{c=1}^k \sum_{j=1}^2 \bar{r}_j^* X_{Tcj}, \quad (4.50)$$

which implies that

$$\text{Bias}(\hat{Y}_T) = \sum_{c=1}^k \sum_{j=1}^2 \text{Bias}(\hat{Y}_{Tcj})$$

But from equation (4.4), $\text{Bias}(\hat{Y}_{Tcj}) = -(N_{cj} - 1)S_{rxcj}$ and that the unbiased estimator for $\text{Bias}(\hat{Y}_{Tcj})$ is

$$\widehat{\text{Bias}}(\hat{Y}_{Tcj}) = -(N_1 - 1)s_{rxcj}$$

which can be expanded as

$$\widehat{\text{Bias}}(\hat{Y}_{Tcj}) = -\frac{N_{cj} - 1}{n_{cj} - 1} \{y_{tcj} - \bar{r}_1^* x_{tcj}\} \quad (4.51)$$

By definition, bias in \hat{Y}_{Tcj} is given as

$$\text{Bias}(\hat{Y}_{Tcj}) = E(\hat{Y}_{Tcj}) - Y_{Tcj}$$

so that the unbiased estimator for Y_{Tcj} is given by

$$\hat{Y}_{Tcj} - \widehat{\text{Bias}}(\hat{Y}_{Tcj})$$

where $\widehat{Y}_{Tcj} = \bar{r}_j^* X_{Tcj}$

That is, the unbiased estimator for Y_{Tcj} can be expressed as

$$\widehat{Y}_{Tcj} = \bar{r}_j^* X_{Tcj} + \frac{N_{cj} - 1}{n_{cj} - 1} \{y_{tcj} - \bar{r}_j^* x_{tcj}\} \quad (4.52)$$

so that summing over all the strata and in both response groups, the estimator for the overall finite population total becomes

$$\widehat{Y}_T = \sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_j^* X_{Tcj} + \frac{N_{cj} - 1}{n_{cj} - 1} \{y_{tcj} - \bar{r}_j^* x_{tcj}\} \right]$$

Hence the derivation.

To study some asymptotic properties of \widehat{Y}_{DC} , the following lemma is considered.

Lemma 4.2: The sample ratio mean \bar{r}^* is unbiased for the population ratio mean \bar{R}^*

Proof. To prove the lemma, we need to show that $E(\bar{r}^*) = \bar{R}^*$.

Now,

$$\begin{aligned} E(\bar{r}^*) &= E\left(\frac{1}{n} \sum_{i=1}^n R_i^*\right) \\ &= \frac{1}{n} \sum_{i=1}^n E(R_i^*) \\ &= \frac{1}{n} n \bar{R}^* \\ &= \bar{R}^* \end{aligned}$$

Hence the proof. □

Using the the proof in Lemma 4.2, it follows that the sample ratio mean for the j^{th} group \bar{r}_j^* is unbiased for population ratio mean for the j^{th} group \bar{R}_j^*

Theorem 4.5: The estimator \widehat{Y}_{DC} is unbiased estimator of the finite population total Y_T

Proof. Here, we need to show that $E(\widehat{Y}_{DC}) = Y_T$.

That is,

$$E\left(\widehat{Y}_{DC}\right) = E\left(\sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_j^* X_{Tcj} + \frac{N_{cj} - 1}{n_{cj} - 1} \{y_{tcj} - \bar{r}_j^* x_{tcj}\}\right]\right) = Y_T$$

where,

$$E\left(\widehat{Y}_{DC}\right) = \sum_{c=1}^k \sum_{j=1}^2 \left[E\left(\bar{r}_j^* X_{Tcj}\right) + \frac{n_{cj}(N_{cj} - 1)}{n_{cj} - 1} \{E(\bar{y}_{cj}) - E(\bar{r}_j^* \bar{x}_{cj})\} \right] \quad (4.53)$$

Using the result in Lemma 4.2 and assuming a large population and consequently a large sample such that $\frac{1}{n_{cj}-1}$ approaches $\frac{1}{n_{cj}}$ and $\frac{1}{N_{cj}-1}$ is so close to $\frac{1}{N_{cj}}$, then equation (4.53) reduces to

$$E\left(\widehat{Y}_{DC}\right) = \sum_{c=1}^k \sum_{j=1}^2 \left[\bar{R}_j^* X_{Tcj} + N_{cj} \{E(\bar{y}_{cj}) - E(\bar{r}_j^* \bar{x}_{cj})\} \right] \quad (4.54)$$

Now, $E(\bar{r}_j^* \bar{x}_{cj}) = E(\bar{r}_j^*) E(\bar{x}_{cj})$ since $Cov(\bar{r}_j^*, \bar{x}_{cj}) = 0$.

Also, under SRSWOR, \bar{y}_{cj} is an unbiased estimator of \bar{Y}_{cj} so that equation

(4.54) can be expressed as

$$\begin{aligned}
E(\widehat{Y}_{DC}) &= \sum_{c=1}^k \sum_{j=1}^2 \left[\overline{R}_j^* X_{Tcj} + N_{cj} \{ \overline{Y}_{cj} - E(\overline{r}_j^*) E(\overline{x}_{cj}) \} \right] \\
&= \sum_{c=1}^k \sum_{j=1}^2 \left[\overline{R}_j^* X_{Tcj} + N_{cj} \{ \overline{Y}_{cj} - \overline{R}_j^* \overline{X}_{cj} \} \right] \\
&= \sum_{c=1}^k \sum_{j=1}^2 \left[\overline{R}_j^* X_{Tcj} + N_{cj} \overline{Y}_{cj} - N_{cj} \overline{R}_j^* \overline{X}_{cj} \right] \\
&= \sum_{c=1}^k \sum_{j=1}^2 \left[\overline{R}_j^* X_{Tcj} + Y_{Tcj} - \overline{R}_j^* X_{Tcj} \right] \\
&= \sum_{c=1}^k \sum_{j=1}^2 Y_{Tcj} \\
&= Y_T
\end{aligned}$$

Hence the proof. □

Using \widehat{Y}_{DC} as an estimator for Y_T , $MSE(\widehat{Y}_{DC})$ is expressed as,

$$\begin{aligned}
MSE(\widehat{Y}_{DC}) &= E[\widehat{Y}_{DC} - Y_T]^2 \\
&= E[\widehat{Y}_{DC} + E(\widehat{Y}_{DC}) - E(\widehat{Y}_{DC}) - Y_T]^2 \\
&= E[\widehat{Y}_{DC} - E(\widehat{Y}_{DC})]^2 + [E(\widehat{Y}_{DC}) - Y_T]^2 \\
&= Var(\widehat{Y}_{DC}) + [Bias(\widehat{Y}_{DC})]^2
\end{aligned}$$

Since \widehat{Y}_{DC} is unbiased, the expression for $MSE(\widehat{Y}_{DC})$ reduces to,

$$MSE(\widehat{Y}_{DC}) = Var(\widehat{Y}_{DC}) \quad (4.55)$$

Theorem 4.6: Under a large sample assumption, variance of \widehat{Y}_{DC} is given as

$$Var(\widehat{Y}_{DC}) = \sum_{c=1}^k \sum_{j=1}^2 \left\{ \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \overline{R}_j^{*2} S_{xcj}^2 - 2\overline{R}_j^* \rho_{cj} S_{xcj} S_{ycj} \right] \right\} \quad (4.56)$$

Proof. To evaluate the expression for $Var(\widehat{Y}_{DC})$, a similar approach as done for $Var(\widehat{Y}_D)$ is used. From equation (4.47)

$$\widehat{Y}_{DC} = \sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_j^* X_{Tcj} + \frac{N_{cj} - 1}{n_{cj} - 1} \{y_{tcj} - \bar{r}_j^* x_{tcj}\} \right]$$

Using the results by Hartley and Ross (1954) and Goodman and Hartley (1958), the general expression for $Var(\widehat{Y}_{DC})$ is

$$Var(\widehat{Y}_{DC}) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_j^{*2} S_{xcj}^2 - 2\bar{R}_j^* S_{xycj} \right. \\ \left. + \frac{1}{n_{cj} - 1} \{S_{r^*j}^2 S_{xcj}^2 + S_{r^*xcj}\} \right]$$

However, $S_{r^*xcj} = 0$, for all $j = 1, 2$ and $c = 1, 2, \dots, k$ so that $Var(\widehat{Y}_{DC})$ reduces to

$$Var(\widehat{Y}_{DC}) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_j^{*2} S_{xcj}^2 - 2\bar{R}_j^* S_{xycj} \right. \\ \left. + \frac{1}{n_{cj} - 1} \{S_{r^*j}^2 S_{xcj}^2\} \right] \quad (4.57)$$

For large samples such that the coefficient $\frac{1}{(n_{cj}-1)}$ asymptotically approaches zero and becomes negligible, the second part of equation (4.57) becomes negligible so that $Var(\widehat{Y}_{DC})$ becomes

$$Var(\widehat{Y}_{DC}) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_j^{*2} S_{xcj}^2 - 2\bar{R}_j^* S_{xycj} \right] \quad (4.58)$$

or equivalently,

$$Var(\widehat{Y}_{DC}) = \sum_{c=1}^k \sum_{j=1}^2 \left\{ \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_j^{*2} S_{xcj}^2 - 2\bar{R}_j^* \rho_{cj} S_{xcj} S_{ycj} \right] \right\} \quad (4.59)$$

Hence the proof. \square

The expression in equation (4.59) implies that $MSE(\widehat{Y}_{DC})$, or equivalently $Var(\widehat{Y}_{DC})$, decreases as the c^{th} stratum sample size in response group j becomes large. Also, the precision of Y_{DC} is improved for a sufficiently large

ρ_j . Therefore, for both separate and combined ratio estimation methods, variances of \widehat{Y}_D and \widehat{Y}_{DC} reduces when the sample size increases and when there is a perfect correlation between Y and X .

Using Corollary 1 under Theorem 4.2, it can be shown that

$$Var(\widehat{Y}_{DC}) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left\{ \frac{1}{N_{cj} - 1} \sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{R}_j X_{cij})^2 \right\} \quad (4.60)$$

To determine whether the estimator \widehat{Y}_{DC} is consistent for estimating Y_T , a sequence of point estimators for finite population total, denoted as $\{\widehat{Y}_{DC}\}$, is considered. By definition, the sequence is said to be weakly consistent for Y_T if \widehat{Y}_{DC} converges in probability to Y_T as the sample size becomes large (Cochran, 1977).

Theorem 4.7: The combined ratio estimator \widehat{Y}_{DC} is a consistent estimator for finite population total Y_T .

Proof. Proofing the consistency of \widehat{Y}_{DC} in estimator Y_T involves showing that as the stratum sample size in response group j , n_{cj} tends to N_{cj} , variance of \widehat{Y}_{DC} diminishes such that for every small positive constant ε , the relation,

$$\lim_{n_{cj} \rightarrow N_{cj}} Pr\{|\widehat{Y}_{DC} - Y_T| > \varepsilon\} = 0, \quad (4.61)$$

holds.

That is, \widehat{Y}_{DC} is consistent if the sample sizes in each stratum and in the two response groups becomes large and gets close to the corresponding stratum population sizes in both response groups, its variance tends to *zero*. Or equivalently, as n_{cj} becomes large and gets very close to N_{cj} , for every $\varepsilon > 0$,

$$\begin{aligned} Pr\{|\widehat{Y}_{DC} - Y_T| > \varepsilon\} &\leq \frac{Var(\widehat{Y}_{DC})}{\varepsilon^2} \\ &= \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_j^{*2} S_{xcj}^2 - 2\bar{R}_j^* S_{xycj} \right] \end{aligned} \quad (4.62)$$

Now, as $n_{cj} \rightarrow N_{cj}$, the limit for $Var(\widehat{Y}_{DC})$ is obtained as

$$\lim_{n_{cj} \rightarrow N_{cj}} \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_j^{*2} S_{xcj}^2 - 2\bar{R}_j^* S_{xyej} \right] = 0 \quad (4.63)$$

Thus, $\widehat{Y}_{DC} \xrightarrow{p} Y_T$, which is the condition for consistency.

Hence the proof □

Assuming a large sample normal distribution in both sampling phase I and phase II, the confidence interval of Y_T is given by

$$\widehat{Y}_{DC} \pm Z_{\frac{\alpha}{2}} \sqrt{Var(\widehat{Y}_{DC})} \quad (4.64)$$

where $Z_{\frac{\alpha}{2}}$ is the co-efficient value $(1 - \frac{\alpha}{2})$, which obtained from the standard normal table at $(1 - \alpha)100\%$ confidence level.

4.4 Comparison of Separate and Combined Ratio Estimators

In this section, we compare the efficiency of the estimators \widehat{Y}_D and \widehat{Y}_{DC} .

Theorem 4.8: Performance of \widehat{Y}_D relative to that of \widehat{Y}_{DC} depends on the absolute difference between the strata mean ratio \bar{R}_{cj} in response group j and the j^{th} response group overall population mean ratio \bar{R}_j^* .

Proof. The expressions for \widehat{Y}_D and \widehat{Y}_{DC} are given as

$$\widehat{Y}_D = \sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_{cj} X_{Tcj} + \frac{N_{cj} - 1}{n_{cj} - 1} \{y_{tcj} - \bar{r}_{cj} x_{tcj}\} \right],$$

and

$$\widehat{Y}_{DC} = \sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_j^* X_{Tcj} + \frac{N_{cj} - 1}{n_{cj} - 1} \{y_{tcj} - \bar{r}_j^* x_{tcj}\} \right]$$

with the corresponding variances (or equivalently, MSE's) expressed as

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_{cj}^2 S_{xcj}^2 - 2\bar{R}_{cj} S_{xyej} + \frac{1}{n_{cj} - 1} [S_{rcj}^2 S_{xcj}^2 + S_{rxcj}] \right] \quad (4.65)$$

and

$$Var(\widehat{Y}_{DC}) = \sum_{c=1}^k \sum_{j=1}^2 \left\{ \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_j^{*2} S_{xcj}^2 - 2\bar{R}_j^* S_{xycj} + \frac{1}{n_{cj} - 1} S_{r^*j}^2 S_{xcj}^2 \right] \right\} \quad (4.66)$$

However, under large sample approximations,

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \left\{ \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_{cj}^2 S_{xcj}^2 - 2\bar{R}_{cj} S_{xycj} \right] \right\} \quad (4.67)$$

and

$$Var(\widehat{Y}_{DC}) = \sum_{c=1}^k \sum_{j=1}^2 \left\{ \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_j^{*2} S_{xcj}^2 - 2\bar{R}_j^* S_{xycj} \right] \right\} \quad (4.68)$$

To compare the variances of \widehat{Y}_D and \widehat{Y}_{DC} , it can be noted from equation (4.67) and equation (4.68) that the expressions $Var(\widehat{Y}_D)$ and $Var(\widehat{Y}_{DC})$ differ only in the form of the ratio estimate \bar{R}_{cj} and \bar{R}_j^* .

Let $\Delta = Var(\widehat{Y}_{DC}) - Var(\widehat{Y}_D)$ so that

$$\begin{aligned} \Delta &= \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[\left(\bar{R}_j^{*2} - \bar{R}_{cj}^2 \right) S_{xcj}^2 - 2 \left(\bar{R}_{cj} - \bar{R}_j^* \right) \rho_{cj} S_{xcj} S_{ycj} \right] \\ &= \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[\left(\bar{R}_j^* - \bar{R}_{cj} \right)^2 S_{xcj}^2 + 2 \left(\bar{R}_j^* - \bar{R}_{cj} \right) \left(\bar{R}_{cj} S_{xcj}^2 - \rho_{cj} S_{xcj} S_{ycj} \right) \right] \end{aligned}$$

It can be observed that the difference Δ depends on the absolute difference between the strata mean ratio \bar{R}_{cj} in response group j and the j^{th} response group overall population mean ratio \bar{R}_j^* . Further, assuming that the regression line of Y on X is linear and passes through the origin in each stratum, the value of $\left(\bar{R}_{cj} S_{xcj}^2 - \rho_{cj} S_{xcj} S_{ycj} \right)$ vanishes. That is,

$$\left(\bar{R}_{cj} S_{xcj}^2 - \rho_{cj} S_{xcj} S_{ycj} \right) \text{ tends to zero,}$$

which implies that,

$$\bar{R}_{cj} = \rho_{cj} \frac{S_{ycj}}{S_{xcj}}.$$

Under this condition, $MSE(\widehat{Y}_{DC}) > MSE(\widehat{Y}_D)$ so that \widehat{Y}_D becomes more precise than \widehat{Y}_{DC} . In this case, the level of precision of \widehat{Y}_D will be improved if $\bar{R}_{cj} \neq \bar{R}_j^*$. However, if $\bar{R}_{cj} = \bar{R}_j^*$, \widehat{Y}_{DC} and \widehat{Y}_D have equal performance.

Hence the proof. □

4.5 Chapter Summary

In this chapter, an unbiased ratio-type estimator for finite population total in stratified random sampling under non-response has successfully been constructed. Both separate and combined ratio estimation methods have been considered and in each case, asymptotic properties of the constructed ratio-type estimators have been studied. A comparison of the constructed separate ratio-type estimator and the usual ratio estimator shows that the constructed estimator is more efficient than the usual ratio estimator. From the optimality conditions, we have noted that the suggested unbiased ratio-type estimator is a best linear unbiased estimator for finite population total when \bar{r}_{cj} is a best linear unbiased estimator for \bar{R}_{cj} ($j = 1, 2$). Also, it has been observed that performance of the constructed unbiased ratio estimator under separate and combined ratio estimation methods depends on the absolute difference between the strata mean ratio \bar{R}_{cj} in response group j and the j^{th} response group overall population mean ratio \bar{R}_j^* .

5. REGRESSION-BASED UNBIASED RATIO ESTIMATORS

5.1 Introduction

In this chapter, construction of regression and multivariate forms of unbiased ratio estimator for finite population total in stratified random sampling under non-response is done. Their performance is compared with known ratio estimators.

5.2 Regression Form of \hat{Y}_D

A regression estimator increases the precision of the estimator of finite population total by utilizing some auxiliary information. In the usual ratio estimation, efficiency of the ratio estimator is improved when the regression line of the response variable Y on the auxiliary variable X passes through the origin such that $Y_i = \beta X_i + e_i$, where β is the regression coefficient and e_i is the error term and $i = 1, 2, \dots, N$ (say). This condition of linear regression between Y and X is not always the case. In such cases, regression estimation becomes the best approach. In this study, a regression estimator is constructed using the suggested estimator \hat{Y}_D and properties studied under the assumption of large populations and consequently large sample sizes.

To improve the estimator of population mean \bar{Y} , a general form of the regression estimator

$$\hat{Y}_R = \bar{y} + \mu(\bar{x} - \bar{X}) \text{ is considered} \quad (5.1)$$

where \bar{x} is an unbiased estimator of population mean \bar{X} of the auxiliary variable X , \bar{y} is the conventional sample mean per unit estimator of Y and μ is any constant suitably chosen such that $Var(\hat{Y}_R)$ is minimum.

By definition, the expression for $Var(\widehat{Y}_R)$ is obtained from

$$\begin{aligned}
 Var(\widehat{Y}_R) &= E \left[\widehat{Y}_R - E(\widehat{Y}_R) \right]^2 \\
 &= \frac{N-n}{nN} \frac{\sum_{i=1}^N [(Y_i - \bar{Y}) + \mu(X_i - \bar{X})]^2}{N-1} \\
 &= \frac{N-n}{nN} (Var(\bar{y}) + \mu^2 Var(\bar{x}) + 2\mu Cov(\bar{x}, \bar{y})) \\
 &= \frac{N-n}{nN} (S_y^2 + \mu^2 S_x^2 + 2\mu S_{xy})
 \end{aligned}$$

Now, using ordinary least square method, the optimum value of μ that minimizes $Var(\widehat{Y}_R)$ is

$$\widehat{\mu}_{opt} = -\frac{Cov(\bar{x}, \bar{y})}{Var(\bar{x})} = -\frac{S_{xy}}{S_x^2} \quad (5.2)$$

Consider the linear model

$$Y = \beta X + e \quad (5.3)$$

where e is an error term that arise due to lack of exact relationship between the auxiliary variable X and the response variable Y . The optimum value of β is obtained by minimizing $\sum e_i^2$ based on n paired observations $(x_i, y_i), i = 1, 2, \dots, n$, which is obtained as

$$\widehat{\beta}_{opt} = \frac{Cov(\bar{x}, \bar{y})}{Var(\bar{x})} = \frac{S_{xy}}{S_x^2} \quad (5.4)$$

Comparing equation (5.2) and equation (5.4) gives

$$\widehat{\mu}_{opt} = -\widehat{\beta}_{opt}$$

so that \widehat{Y}_R can now be expressed as

$$\widehat{Y}_R = \bar{y} + \beta(\bar{X} - \bar{x}) \quad (5.5)$$

with a corresponding variance expressed as

$$Var(\widehat{Y}_R) = \frac{N-n}{nN} (S_y^2 + \beta^2 S_x^2 - 2\beta S_{xy}) \quad (5.6)$$

which simplifies to

$$Var(\widehat{Y}_R)_{min} = \frac{N-n}{nN} (1 - \rho_{xy}^2) S_Y^2 \quad (5.7)$$

where ρ_{XY} is the correlation coefficient between X and Y .

Unbiased sample estimates for $Var(\widehat{Y}_R)$ and $Var(\widehat{Y}_R)_{min}$ are therefore given as

$$v(\widehat{Y}_R) = \frac{N-n}{nN} (s_y^2 + \beta^2 s_x^2 - 2\beta s_{xy})$$

and

$$v(\widehat{Y}_R)_{min} = \frac{N-n}{nN} (1 - \rho_{xy}^2) s_y^2$$

Now, to obtain the regression estimator for population total under SRSWOR, equation (5.5) is multiplied by N and its corresponding minimum variance is obtained by multiplying equation (5.7) by N^2 .

As previously noted in stratified random sampling with k strata, that we can either have separate regression estimator or combined regression estimator for finite population total depending on whether we use the form \widehat{Y}_D or \widehat{Y}_{DC} in the estimation procedure. Separate regression estimator for finite population total can be expressed as

$$\widehat{Y}_{R(st)} = \sum_{c=1}^k N_c \widehat{Y}_{Rc} \quad (5.8)$$

where $\widehat{Y}_{Rc} = \bar{y}_c + \beta_c(\bar{X}_c - \bar{x}_c)$ is the regression mean estimate in each stratum.

To obtain a combined regression estimator, all sample information in all strata is combined first and then implemented in the regression estimator.

Let \widehat{Y}_{RC} denote the combined regression estimator for population mean, then by defining $\bar{y}_{st} = \sum_{c=1}^k \frac{N_c}{N} \widehat{Y}_{Rc}$ and $\bar{x}_{st} = \sum_{c=1}^k \frac{N_c}{N} \bar{X}_{Rc}$, \widehat{Y}_{RC} can be obtained using

$$\widehat{Y}_{RC} = \bar{y}_{st} + \beta(\bar{X} - \bar{x}_{st}) \quad (5.9)$$

Even though several studies have been done on regression estimation method, focus has been on using complete data to estimate population mean. In this study, separate regression estimation is considered and a regression estimator is constructed under non-response in the response variable. A regression

form of the estimator \widehat{Y}_D , denoted as \widehat{Y}_{DR} , is therefore suggested as

$$\widehat{Y}_{DR} = \sum_{j=1}^2 \sum_{c=1}^k N_{cj} \left[\widehat{Y}_{cj}^* + \beta_{cj} (\overline{X}_{cj} - \overline{x}_{cj}) \right] \quad (5.10)$$

where $\widehat{Y}_{cj}^* = \frac{\widehat{Y}_{Tcj}}{N_{cj}}$.

For c^{th} stratum and in j^{th} response group, \widehat{Y}_{DRcj} is expressed as

$$\widehat{Y}_{DRcj} = N_{cj} \left[\widehat{Y}_{cj}^* + \beta_{cj} (\overline{X}_{cj} - \overline{x}_{cj}) \right] \quad (5.11)$$

To study some properties of \widehat{Y}_{DR} , an assumption that the regression coefficient is known in prior as β_{0cj} (say) is made, so that \widehat{Y}_{DR} is expressed as

$$\widehat{Y}_{DR} = \sum_{j=1}^2 \sum_{c=1}^k N_{cj} \left[\widehat{Y}_{cj}^* + \beta_{0cj} (\overline{X}_{cj} - \overline{x}_{cj}) \right]$$

Theorem 5.1: \widehat{Y}_{DR} is an unbiased estimator for finite population total Y_T .

Proof. This proof involves showing that $E(\widehat{Y}_{DR}) = Y_T$.

Consider a random sample $(x_i, y_i), i = 1, 2, \dots, n$ drawn by SRSWOR. Then $E(\widehat{Y}_{DR})$ is obtained as follows

$$\begin{aligned} E(\widehat{Y}_{DR}) &= \sum_{j=1}^2 \sum_{c=1}^k N_{cj} E \left[\widehat{Y}_{cj}^* + \beta_{0cj} (\overline{X}_{cj} - \overline{x}_{cj}) \right] \\ &= \sum_{j=1}^2 \sum_{c=1}^k N_{cj} \left[\overline{Y}_{cj}^* + \beta_{0cj} (\overline{X}_{cj} - E(\overline{x}_{cj})) \right] \\ &= \sum_{j=1}^2 \sum_{c=1}^k N_{cj} \left[\overline{Y}_{cj}^* + \beta_{0cj} (\overline{X}_{cj} - \overline{X}_{cj}) \right] \\ &= \sum_{j=1}^2 \sum_{c=1}^k N_{cj} \overline{Y}_{cj}^* \\ &= Y_T \end{aligned}$$

Hence the proof. □

To obtain the expression for $Var(\widehat{Y}_{DR})$, a case of an arbitrary value for the regression coefficient is considered.

Theorem 5.2: For a known value of the regression coefficient, the expression of $Var(\widehat{Y}_{DR})$ is given as

$$\sum_{j=1}^2 \sum_{c=1}^k \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^{2*} + \beta_{0cj}^2 S_{xcj}^2 - 2\beta_{0cj} S_{xycj} \right] \quad (5.12)$$

where S_{ycj}^{2*} is $Var(\widehat{Y}_{cj}^*)$.

Proof. Here, we obtain the expression of $Var(\widehat{Y}_{DR})$ for a known regression coefficient and we proceed as follows,

$$\begin{aligned} Var(\widehat{Y}_{DR}) &= E \left[\widehat{Y}_{DR} - E(\widehat{Y}_{DR}) \right]^2 \\ &= E \left[\sum_{j=1}^2 \sum_{c=1}^k N_{cj} \left[\widehat{Y}_{cj}^* + \beta_{0cj} (\bar{X}_{cj} - \bar{x}_{cj}) \right] - Y_T \right]^2 \\ &= E \left[\sum_{j=1}^2 \sum_{c=1}^k N_{cj} \left[\widehat{Y}_{cj}^* + \beta_{0cj} (\bar{X}_{cj} - \bar{x}_{cj}) \right] - \sum_{j=1}^2 \sum_{c=1}^k N_{cj} \bar{Y}_{cj} \right]^2 \\ &= \sum_{j=1}^2 \sum_{c=1}^k N_{cj}^2 E \left[\left(\widehat{Y}_{cj}^* - \bar{Y}_{cj} \right) - \beta_{0cj} (\bar{x}_{cj} - \bar{X}_{cj}) \right]^2 \\ &= \sum_{j=1}^2 \sum_{c=1}^k N_{cj}^2 \left[E \left(\widehat{Y}_{cj}^* - \bar{Y}_{cj} \right)^2 + \beta_{0cj}^2 E (\bar{x}_{cj} - \bar{X}_{cj})^2 - 2\beta_{0cj} S_{xycj} \right] \\ &= \sum_{j=1}^2 \sum_{c=1}^k \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^{2*} + \beta_{0cj}^2 S_{xcj}^2 - 2\beta_{0cj} S_{xycj} \right] \end{aligned}$$

where, $S_{xycj} = \beta_{0cj} E (\bar{x}_{cj} - \bar{X}_{cj}) \left(\widehat{Y}_{cj}^* - \bar{Y}_{cj} \right)$

Hence the proof. \square

As n_{cj} becomes large and tends close to N_{cj} , the expression for $Var(\widehat{Y}_{DR})$ tends to zero. That is,

$$\lim_{n_{cj} \rightarrow N_{cj}} Var(\widehat{Y}_{DR}) \rightarrow 0 \quad (5.13)$$

This is an indication that \widehat{Y}_{DR} is not only unbiased, but also a consistent estimator for finite population total.

Corollary 5.1 An unbiased sample estimate of variance of \widehat{Y}_{DR} is thus given as

$$Var(\widehat{Y}_{DR}) = \sum_{j=1}^2 \sum_{c=1}^k \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}(n_{cj} - 1)} \left[s_{ycj}^{2*} + \beta_{0cj}^2 s_{xcj}^2 - 2\beta_{0cj} s_{xycj} \right] \quad (5.14)$$

Now, for the j^{th} response group in stratum c , $Var(\widehat{Y}_{DRcj})$ is obtained from the expression

$$Var(\widehat{Y}_{DRcj}) = \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^{2*} + \beta_{0cj}^2 S_{xcj}^2 - 2\beta_{0cj} S_{xycj} \right] \quad (5.15)$$

The problem now is how to obtain the value of β_{0cj} that minimizes $Var(\widehat{Y}_{DRcj})$ in all the strata.

Theorem 5.3: The optimal value of β_{0cj} that minimizes $Var(\widehat{Y}_{DRcj})$ is given by

$$\beta_{0cj} = B_{cj} = \frac{S_{cjxy}}{S_{cjx}^2} = \frac{\sum_{i=1}^{N_{cj}} (Y_{cij} - \bar{Y}_{cj}) (X_{cij} - \bar{X}_{cj})}{\sum_{i=1}^{N_{cj}} (X_{cij} - \bar{X}_{cj})^2} \quad (5.16)$$

to give a minimum variance expressed as

$$Var(\widehat{Y}_{DRcj})_{min} = \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} S_{ycj}^{2*} \left[1 - \rho_{cj}^2 \right] \quad (5.17)$$

where B_{cj} is the population regression coefficient of Y on X and ρ_{cj} is the population correlation coefficient between X and Y , both in the j^{th} response group in stratum c .

Proof. To proof that the optimal value of β_{0cj} is expressed as given in equation (5.16) and $Var(\widehat{Y}_{DRcj})_{min}$ is obtained using the expression in equation (5.17), the value of B_{cj} can theoretically be pre-assigned since it does not depend on the characteristic of any sample drawn.

Now, in equation (5.15), let

$$\beta_{0cj} = B_{cj} + q = \frac{S_{xycj}}{S_{xcj}^2} + q, \text{ where } q \text{ is any constant} \quad (5.18)$$

to obtain

$$\begin{aligned} Var(\widehat{Y}_{DRcj}) &= \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^{2*} + \left(\frac{S_{xycj}}{S_{xcj}^2} + q \right)^2 S_{xcj}^2 - 2 \left(\frac{S_{xycj}}{S_{xcj}^2} + q \right) S_{xycj} \right] \\ &= \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^{2*} + \left(\frac{S_{xycj}^2}{S_{xcj}^4} + 2q \frac{S_{xycj}}{S_{xcj}^2} + q^2 \right) S_{xcj}^2 - 2 \left(\frac{S_{xycj}}{S_{xcj}^2} + q \right) S_{xycj} \right] \\ &= \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[\left(S_{ycj}^{2*} - \frac{S_{xycj}^2}{S_{xcj}^2} \right) + q^2 S_{xcj}^2 \right] \end{aligned} \quad (5.19)$$

Clearly, equation (5.19) is minimum when $q = 0$. Under this condition,

$$Var(\widehat{Y}_{DRcj})_{min} = \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[\left(S_{ycj}^{2*} - \frac{S_{xycj}^2}{S_{xcj}^2} \right) \right] \quad (5.20)$$

But $S_{xycj} = \rho_{xycj} S_{xcj} S_{ycj}^*$ then,

$$Var(\widehat{Y}_{DRcj})_{min} = \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} S_{ycj}^{2*} [1 - \rho_{cj}^2] \quad (5.21)$$

Hence the proof. \square

Therefore, for the entire population, minimum variance of \widehat{Y}_{DR} is expressed as

$$Var(\widehat{Y}_{DR})_{min} = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} S_{ycj}^{2*} [1 - \rho_{cj}^2] \quad (5.22)$$

To examine the departure of β_{0cj} from B_{cj} without any substantial loss of precision, equations (5.19) and (4.46) are used such that

$$\begin{aligned} Var(\widehat{Y}_{DRcj}) &= \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} [S_{ycj}^{2*} (1 - \rho_{cj}^2) + (\beta_{0cj} - B_{cj})^2 S_{xcj}^2] \\ &= Var(\widehat{Y}_{DR})_{min} \left[1 + \frac{(\beta_{0cj} - B_{cj})^2 S_{xcj}^2}{S_{ycj}^{2*} (1 - \rho_{cj}^2)} \right] \end{aligned} \quad (5.23)$$

But from $B_{cj} = \frac{S_{xycj}^*}{S_{xcj}^2}$, equation (5.23) can be simplified as follows

$$\begin{aligned}
Var(\widehat{Y}_{DRcj}) &= Var(\widehat{Y}_{DR})_{min} \left[1 + \frac{(\beta_{0cj} - B_{cj})^2}{B_{cj}^2} \frac{S_{xcj}^2 B_{cj}^2}{S_{ycj}^{2*} (1 - \rho_{cj}^2)} \right] \\
&= Var(\widehat{Y}_{DR})_{min} \left[1 + \left(\frac{\beta_{0cj}}{B_{cj}} - 1 \right)^2 \frac{S_{xcj}^2}{S_{ycj}^{2*} (1 - \rho_{cj}^2)} \frac{S_{xycj}^2}{S_{xcj}^4} \right] \\
&= Var(\widehat{Y}_{DR})_{min} \left[1 + \left(\frac{\beta_{0cj}}{B_{cj}} - 1 \right)^2 \frac{S_{xycj}^2}{S_{ycj}^{2*} S_{xcj}^2 (1 - \rho_{cj}^2)} \right] \\
&= Var(\widehat{Y}_{DR})_{min} \left[1 + \left(\frac{\beta_{0cj}}{B_{cj}} - 1 \right)^2 \frac{\rho_{cj}^2}{(1 - \rho_{cj}^2)} \right]
\end{aligned} \tag{5.24}$$

Now, using least square method, the value of β_{cj} that minimizes $Var(\widehat{Y}_{DRcj})$ is obtained by differentiating equation (5.15) with respect to β_{cj} and equating to zero and this procedure gives

$$\widehat{\beta}_{cj(opt)} = \frac{S_{xycj}}{S_{xcj}^2} = \frac{\rho_{cj} S_{xcj} S_{ycj}^*}{S_{xcj}^2} \tag{5.25}$$

so that substituting the expression for $\beta_{cj(opt)}$ in equation (5.15) simplifies to

$$Var(\widehat{Y}_{DRcj})_{min} = \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} S_{ycj}^{2*} [1 - \rho_{cj}^2].$$

Theorem 5.4: If β_{cj} is the least square estimate of B_{cj} and

$$\widehat{Y}_{DRcj} = N_{cj} \left[\widehat{Y}_{cj}^* + \beta_{cj} (\overline{X}_{cj} - \bar{x}_{cj}) \right], \tag{5.26}$$

then using simple random samples of size n_{cj} , where n_{cj} is large,

$$Var(\widehat{Y}_{DRcj}) = \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} S_{ycj}^{2*} [1 - \rho_{cj}^2]$$

Proof. By definition, the sampling error of an estimator $\widehat{\theta}$ for a parameter θ is obtained from the quantity $\widehat{\theta} - \theta$ (Cochran, 1977). That is, the sampling error of \widehat{Y}_{DRcj} is obtained from the quantity $\widehat{Y}_{DRcj} - Y_{Tcj}$.

Now,

$$\begin{aligned}
\widehat{Y}_{DRcj} - Y_{Tcj} &= N_{cj} \left[\overline{Y}_{cj}^* + \beta_{cj} (\overline{X}_{cj} - \overline{x}_{cj}) \right] - Y_{Tcj} \\
&= N_{cj} \left[\overline{Y}_{cj}^* + \beta_{cj} (\overline{X}_{cj} - \overline{x}_{cj}) - \overline{Y}_{cj} \right] \\
&= N_{cj} \left[\overline{Y}_{cj}^* - \overline{Y}_{cj} + \beta_{cj} (\overline{X}_{cj} - \overline{x}_{cj}) \right]
\end{aligned}$$

For approximations, we substitute β_{cj} with B_{cj} in equation (5.26) to get

$$\widehat{Y}_{DRcj} = N_{cj} \left[\overline{Y}_{cj}^* + B_{cj} (\overline{X}_{cj} - \overline{x}_{cj}) \right] \quad (5.27)$$

In this approximation, the error committed is $(B_{cj} - \beta_{cj}) (\overline{X}_{cj} - \overline{x}_{cj})$. In a simple random sample of size n_{cj} , this quantity is of size $\frac{1}{n_{cj}}$ since $(\overline{X}_{cj} - \overline{x}_{cj})$ and $(B_{cj} - \beta_{cj})$ are both of order $\frac{1}{\sqrt{n_{cj}}}$. But in SRSWOR, the error in the sample mean of the variate $(y_{cij} - Bx_{cij})$ is the same as the sampling error in \widehat{Y}_{DRcj} . Hence, the sampling error in \widehat{Y}_{DRcj} is of order $\frac{1}{\sqrt{n_{cj}}}$ and consequently, the leading term in $E \left(\widehat{Y}_{DRcj} - Y_{Tcj} \right)^2$ is $Var(\widehat{Y}_{DRcj})$.

That is, for large samples,

$$E \left(\widehat{Y}_{DRcj} - Y_{Tcj} \right)^2 = Var(\widehat{Y}_{DRcj}) = \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} S_{y_{cj}}^{2*} \left[1 - \rho_{cj}^2 \right] \quad (5.28)$$

Hence the proof. □

Given some preassigned β_{0cj} , the departure from $\beta_{cj(opt)}$ is evaluated as follows,

Theorem 5.5: Given some pre-assigned β_{0cj} and $\beta_{cj(opt)}$ obtained from some sample data, $Var(\widehat{Y}_{DR})$ depends on the difference between β_{0cj} and $\beta_{cj(opt)}$.

Proof. In this proof, $Var(\widehat{Y}_{DR})$ is shown to depend on the departure of β_{0cj} from $\beta_{cj(opt)}$.

From equation (5.25), $\beta_{cj(opt)} = \frac{S_{xycj}}{S_{xcj}^2} = \frac{\rho_{cj}S_{ycj}}{S_{xcj}}$ so that

$$\begin{aligned} Var(Y_{DR}) &= \sum_{j=1}^2 \sum_{c=1}^k \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^{2*} + \beta_{0cj}^2 S_{xcj}^2 - 2\beta_{0cj} S_{xycj} \right] \\ &= \sum_{j=1}^2 \sum_{c=1}^k \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^{2*} + \beta_{0cj}^2 S_{xcj}^2 - 2\beta_{0cj} \rho_{cj} S_{xcj} S_{ycj} \right] \\ &= \sum_{j=1}^2 \sum_{c=1}^k \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^{2*} + \beta_{0cj}^2 S_{xcj}^2 - 2\beta_{0cj} \rho_{cj} S_{xcj} S_{ycj} - \rho_{cj}^2 S_{ycj}^{2*} + \rho_{cj}^2 S_{ycj}^{2*} \right] \end{aligned}$$

But from $\beta_{cj(opt)} = \frac{\rho_{cj}S_{ycj}^*}{S_{xcj}}$, $\rho_{cj} S_{xcj} S_{ycj}^* = S_{xcj}^2 \beta_{cj(opt)}$ so that $Var(Y_{DR})$ can be expressed as

$$\begin{aligned} Var(Y_{DR}) &= \sum_{j=1}^2 \sum_{c=1}^k \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[(1 - \rho_{cj}^2) S_{ycj}^{2*} + \beta_{0cj}^2 S_{xcj}^2 - 2\beta_{0cj} S_{xcj}^2 \beta_{cj(opt)} + \beta_{cj(opt)}^2 S_{xcj}^2 \right] \\ &= \sum_{j=1}^2 \sum_{c=1}^k \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[(1 - \rho_{cj}^2) S_{ycj}^{2*} + (\beta_{0cj} - \beta_{cj(opt)})^2 S_{xcj}^2 \right] \end{aligned} \quad (5.29)$$

From equation (5.29), it can be noted that as the difference between β_{0cj} and $\beta_{cj(opt)}$ increases, variance of \hat{Y}_{DR} increases.

Hence the proof. □

5.3 Efficiency of \hat{Y}_{DR}

In this section, efficiency of \hat{Y}_{DR} relative to that of the mean per unit estimator $\hat{Y}_{T(SRS)}$ and the usual ratio estimator \hat{Y}_{TR} is evaluated. This is done under stratified random sampling. Also, efficiency \hat{Y}_{DR} relative to that of \hat{Y}_D is compared.

Using stratum c and response group j , the MSE's under the assumption of large samples are expressed as

$$MSE(\hat{Y}_{DRcj}) = \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[(1 - \rho_{cj}^2) S_{ycj}^{2*} \right]$$

$$MSE(\widehat{Y}_{Dcj}) = \frac{N_{cj}(N_{cj}-n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \bar{R}_{cj}^2 S_{xcj}^2 - 2\bar{R}_{cj}\rho_{cj}S_{xcj}S_{ycj} \right]$$

$$MSE(\widehat{Y}_{TRcj}) = \frac{N_{cj}(N_{cj}-n_{cj})}{n_{cj}} \left[S_{ycj}^2 + R_{cj}^2 S_{xcj}^2 - 2R_{cj}\rho_{cj}S_{xcj}S_{ycj} \right]$$

$$MSE(\widehat{Y}_{Tcj})_{SRS} = \frac{N_{cj}(N_{cj}-n_{cj})}{n_{cj}} S_{ycj}^2$$

Comparing $MSE(\widehat{Y}_{DRcj})$ and $MSE(\widehat{Y}_{Tcj})_{SRS}$, it can be observed that $MSE(\widehat{Y}_{DRcj}) = (1 - \rho_{cj}^2) MSE(\widehat{Y}_{Tcj})_{SRS}$.

But $0 < \rho_{cj}^2 < 1$, which implies that $MSE(\widehat{Y}_{DRcj}) < MSE(\widehat{Y}_{Tcj})_{SRS}$. Hence, \widehat{Y}_{DRcj} is more precise than $\widehat{Y}_{T(SRS)}$ and consequently, \widehat{Y}_{DR} performs better than $\widehat{Y}_{T(SRS)}$.

Comparison of \widehat{Y}_{DRcj} and \widehat{Y}_{TRcj} shows that \widehat{Y}_{DRcj} gives precise estimates than \widehat{Y}_{TRcj} only if $MSE(\widehat{Y}_{DRcj}) < MSE(\widehat{Y}_{TRcj})$

$$\text{or if } (1 - \rho_{cj}^2) S_{ycj}^2 \leq S_{ycj}^2 + R_{cj}^2 S_{xcj}^2 - 2R_{cj}\rho_{cj}S_{xcj}S_{ycj}$$

$$\text{or if } (R_{cj}S_{xcj} - \rho_{cj}S_{ycj})^2 \geq 0$$

which is always true for all $R_{cj}, S_{xcj}, \rho_{cj}, S_{ycj}$

Hence, \widehat{Y}_{DRcj} performs better than \widehat{Y}_{TRcj} and consequently \widehat{Y}_{DR} is better than \widehat{Y}_{TR} .

For \widehat{Y}_{DRcj} and \widehat{Y}_{Dcj} , \widehat{Y}_{DRcj} is more efficient than \widehat{Y}_{Dcj} only if $MSE(\widehat{Y}_{DRcj}) < MSE(\widehat{Y}_{Dcj})$

$$\text{or if } (1 - \rho_{cj}^2) S_{ycj}^2 \leq S_{ycj}^2 + \bar{R}_{cj}^2 S_{xcj}^2 - 2\bar{R}_{cj}\rho_{cj}S_{xcj}S_{ycj}$$

$$\text{or if } (\bar{R}_{cj}S_{xcj} - \rho_{cj}S_{ycj})^2 \geq 0$$

which is always true for all $\bar{R}_{cj}, S_{xcj}, \rho_{cj}, S_{ycj}$

Hence, \widehat{Y}_{DRcj} is a better estimator than \widehat{Y}_{TRcj} , consequently \widehat{Y}_{DR} is more efficient than \widehat{Y}_{TR} .

Clearly, from the comparisons of \widehat{Y}_{DR} and other ratio estimators, we have observed that \widehat{Y}_{DR} performs better than other ratio estimators.

5.4 Multivariate Form of \widehat{Y}_D

In multivariate ratio estimation, an estimator is obtained using a p-component auxiliary random vector. In this section, therefore, we construct a multivariate form of the unbiased ratio estimator in equation (4.1). That is, a multi-auxiliary variables $\underline{X}_i, i = 1, 2, \dots, N$ is considered.

In this section, the general multivariate ratio form suggested by Olkin (1958) and the multivariate ratio form in stratified random sampling suggested by Ngesa et al. (2012) is used. Olkin (1958) suggested a multivariate ratio estimator for population total under simple random sampling scheme as

$$\widehat{Y}_{MR} = \sum_{l=1}^p W_l \frac{\bar{y}}{\bar{x}_l} X_l = \sum_{l=1}^p W_l r_l X_l \quad (5.30)$$

where W_l 's are weights with a linear constraint that $\sum_{l=1}^p W_l = 1$

While extending the work of Olkin (1958), Ngesa et al. (2012) considered a stratified random sampling scheme with varying weights in each stratum and defined a multivariate ratio estimator for finite population total using two auxiliary variables as

$$\widehat{Y}_{MRE} = \sum_{c=1}^k \widehat{Y}_{MRc} \quad (5.31)$$

where \widehat{Y}_{MRc} is the stratum estimator for stratum finite population total in stratum c and is computed from the relation

$$\widehat{Y}_{MRc} = W_{c1} \widehat{Y}_{Tc1} + W_{c2} \widehat{Y}_{Tc2}$$

subject to the condition that $\sum_{l=1}^p W_{cl} = 1$ and that $\widehat{Y}_{Tcl} = \frac{\bar{y}}{\bar{x}_l} X_l, c = 1, 2, \dots, k$.

Using simulated data, Ngesa et al. (2012) observed that their proposed estimator in equation (5.31) had a smaller bias compared to Olkin's. Multivariate ratio estimator constructed by Olkin (1958) and Ngesa et al. (2012) were, however, constructed using a biased ratio estimator. Moreover, the estimators \widehat{Y}_{MR} and \widehat{Y}_{MRE} do not address the problem of non-response. To construct a multivariate form of the estimator Y_D , X_{cijl} shall be used to denote i^{th} observation for the l^{th} auxiliary variable in partition j within stratum c where $i = 1, 2, \dots, N_{cj}$, $j = 1, 2$ and $c = 1, 2, \dots, k$. Using Y_{cij} and X_{cijl} , a multivariate ratio estimator is defined as

$$\widehat{Y}_{MD} = \sum_{c=1}^k \sum_{j=1}^2 \sum_{l=1}^p W_{cjl} \left[\bar{r}_{cjl} X_{Tcjl} + \frac{N_{cj} - 1}{n_{cj} - 1} (y_{tcj} - \bar{r}_{cjl} x_{tcjl}) \right] \quad (5.32)$$

where \bar{r}_{cjl} is the mean observation ration for the l^{th} auxiliary variable in partition j within stratum c and X_{Tcjl} is the the l^{th} auxiliary variable population total.

Now, equation (5.32) can equivalently be expressed as

$$\widehat{Y}_{MD} = \sum_{c=1}^k \sum_{j=1}^2 \left[W_{cj1} \widehat{Y}_{Tcj1} + W_{cj2} \widehat{Y}_{Tcj2+}, \dots, + W_{cjp} \widehat{Y}_{Tcjp} \right] \quad (5.33)$$

so that, from equation (5.33), the expression for the multivariate unbiased ratio estimator in particular stratum, c , say, is given as

$$\widehat{Y}_{MDc} = \sum_{j=1}^2 \left(W_{cj1} \widehat{Y}_{Tcj1} + W_{cj2} \widehat{Y}_{Tcj2+}, \dots, + W_{cjp} \widehat{Y}_{Tcjp} \right) \text{ for } j = 1, 2 \quad (5.34)$$

Equation (5.32) is the general multivariate form of the estimator (4.2) under the linear constraint that $\sum_{l=1}^p W_{cjl} = 1$.

In the subsequent steps, we shall consider only stratum c (say) and j^{th} partition, where the multivariate unbiased ratio estimator for finite population total in stratified random sampling under non-response is expressed as

$$\widehat{Y}_{MDcj} = W_{cj1} \widehat{Y}_{Tcj1} + W_{cj2} \widehat{Y}_{Tcj2+}, \dots, + W_{cjp} \widehat{Y}_{Tcjp} \quad (5.35)$$

Next, we wish to obtain variance of the multivariate form of the proposed estimator. Using \widehat{Y}_{MDcj} as the unbiased multivariate ratio estimator for Y_{Tcj} . In this property, two cases of the auxiliary variable are considered.

Case I: When $p = 2$,

Theorem 5.6: For a two component auxiliary random vector, variance of the multivariate unbiased ratio estimator \widehat{Y}_{MD} is given as

$$V(\widehat{Y}_{MD}) = \sum_{c=1}^k \sum_{j=1}^2 \left[W_{cj1}^2 V_{cj1} + 2W_{cj1}W_{cj2}V_{cj12} + W_{cj2}^2 V_{cj2} \right] \quad (5.36)$$

Proof. For $p = 2$, equation (5.34) reduces to

$$\widehat{Y}_{MDcj} = \sum_{j=1}^2 \left(W_{cj1} \widehat{Y}_{Dcj1} + W_{cj2} \widehat{Y}_{Dcj2} \right) \quad (5.37)$$

so that for $j = 1$, we have

$$\widehat{Y}_{MDc1} = W_{c11} \widehat{Y}_{Dc11} + W_{c12} \widehat{Y}_{Dc12} \quad (5.38)$$

Now, for any c^{th} stratum, say, the population total Y_{Tc} is obtained from $Y_{Tc} = \sum_{j=1}^2 Y_{Tcj}$ so that if we subtract Y_{Tc1} from both sides of equation (5.38), we obtain

$$\widehat{Y}_{MDc1} - Y_{Tc1} = W_{c11} \widehat{Y}_{Dc11} + W_{c12} \widehat{Y}_{Dc12} - Y_{Tc1} \quad (5.39)$$

But in each stratum and in each response group, $W_{cj1} + W_{cj2} = 1$, which implies that

$$Y_{Tc1} = (W_{c11} + W_{c12})Y_{Tc1} \quad (5.40)$$

Replacing equation (5.40) to the right hand side of equation (5.39) gives

$$\begin{aligned} \widehat{Y}_{MDc1} - Y_{Tc1} &= W_{c11} \widehat{Y}_{Dc11} + W_{c12} \widehat{Y}_{Dc12} - (W_{c11} + W_{c12})Y_{Tc1} \\ &= W_{c11}(\widehat{Y}_{Dc11} - Y_{Tc1}) + W_{c12}(\widehat{Y}_{Dc12} - Y_{Tc1}) \end{aligned} \quad (5.41)$$

Squaring both sides of equation (5.41) and taking expectation results to

$$V(\widehat{Y}_{MDc1}) = W_{c11}^2 V(\widehat{Y}_{Dc11}) + 2W_{c11}W_{c12}Cov(\widehat{Y}_{Dc11}, \widehat{Y}_{Dc12}) + W_{c12}^2 V(\widehat{Y}_{Dc12}) \quad (5.42)$$

By letting $V_{c11} = \text{Var}(\widehat{Y}_{Dc11})$, $V_{c12} = \text{Var}(\widehat{Y}_{Dc12})$ and $V_{c112} = \text{Cov}(\widehat{Y}_{Dc11}, \widehat{Y}_{Dc12})$, equation (5.42) can be written as

$$V(\widehat{Y}_{MDc1}) = W_{c11}^2 V_{c11} + 2W_{c11}W_{c12}V_{c112} + W_{c12}^2 V_{c12} \quad (5.43)$$

In matrix form,

$$\begin{aligned} V(\widehat{Y}_{MDc1}) &= V\left(W_{c11}\widehat{Y}_{Dc11} + W_{c12}\widehat{Y}_{Dc12}\right) \\ &= V\left(\begin{bmatrix} W_{c11} & W_{c12} \end{bmatrix} \begin{bmatrix} \widehat{Y}_{Dc11} \\ \widehat{Y}_{Dc12} \end{bmatrix}\right) \\ &= \begin{bmatrix} W_{c11} & W_{c12} \end{bmatrix} \begin{bmatrix} V_{c11} & V_{c112} \\ V_{c112} & V_{c12} \end{bmatrix} \begin{bmatrix} W_{c11} \\ W_{c12} \end{bmatrix} \\ &= \underline{W_{c1}}' \Sigma \underline{W_{c1}} \end{aligned} \quad (5.44)$$

where Σ is variance-covariance matrix.

Therefore, for the entire population, variance of \widehat{Y}_{MD} is expressed as

$$V(\widehat{Y}_{MD}) = \sum_{c=1}^k \sum_{j=1}^2 \left[W_{cj1}^2 V_{cj1} + 2W_{cj1}W_{cj2}V_{cj12} + W_{cj2}^2 V_{cj2} \right] \quad (5.45)$$

Hence the proof. \square

Case II: When $p \geq 3$

Theorem 5.7: For $p \geq 3$, variance of \widehat{Y}_{MD} is obtained using the expression

$$V(\widehat{Y}_{MD}) = \sum_{c=1}^k \sum_{j=1}^2 \sum_{l=1}^p W_{cjl}^2 V_{cjl} + 2 \sum_{c=1}^k \sum_{j=1}^2 \sum_{l=1}^{p-1} W_{cjl}W_{cjl'} V_{cjl'l'} \quad (5.46)$$

Proof. For $p \geq 3$, the multivariate estimator in stratum c takes the general form

$$\widehat{Y}_{MDcj} = W_{cj1}\widehat{Y}_{Dcj1} + W_{cj2}\widehat{Y}_{Dcj2} + \dots + W_{cjp}\widehat{Y}_{Dcjp} \quad (5.47)$$

For a particular response group $j = 1$ (say), the multivariate estimator becomes

$$\widehat{Y}_{MDc1} = W_{c11}\widehat{Y}_{Dc11} + W_{c12}\widehat{Y}_{Dc12} + \dots + W_{c1p}\widehat{Y}_{Dc1p} \quad (5.48)$$

Now, subtracting \widehat{Y}_{Tc1} from equation (5.48) gives

$$\widehat{Y}_{MDc1} - Y_{Tc1} = W_{c11}\widehat{Y}_{Dc11} + W_{c12}Y_{Dc12} + \dots + W_{c1p}\widehat{Y}_{Dc1p} - Y_{Tc1} \quad (5.49)$$

Substituting the linear constraint $\sum_{l=1}^p W_{c1l} = 1$ in equation (5.49) and rearranging gives

$$\widehat{Y}_{MDc1} - Y_{Tc1} = W_{c11}(\widehat{Y}_{Dc11} - Y_{Tc1}) + W_{c12}(\widehat{Y}_{Dc12} - Y_{Tc1}) + \dots + W_{c1p}(\widehat{Y}_{Dc1p} - Y_{Tc1}) \quad (5.50)$$

Squaring both sides of equation (5.50) and taking expectation simplifies to

$$V(\widehat{Y}_{MDc1}) \sum_{l=1}^p W_{c1l}^2 V_{c1l} + 2 \sum_{l=1}^{p-1} W_{c1l} W_{c1l'} V_{c1ll'} \quad (5.51)$$

where $V_{c1l} = \text{Variance}(\widehat{Y}_{Dc1l})$, $V_{c1ll'} = \text{Covariance}(\widehat{Y}_{Dc1l}, \widehat{Y}_{Dc1l'})$ and $l' \neq l$

For the entire population, the expression for variance of \widehat{Y}_{MD} takes the form

$$V(\widehat{Y}_{MD}) = \sum_{c=1}^k \sum_{j=1}^2 \sum_{l=1}^p W_{cjl}^2 V_{cjl} + 2 \sum_{c=1}^k \sum_{j=1}^2 \sum_{l=1}^{p-1} W_{cjl} W_{cjl'} V_{cjll'} \quad (5.52)$$

Hence the proof. \square

The problem now is obtaining the expressions for the values of W_{cj1} and W_{cj2} that minimize $V(\widehat{Y}_{MD})$ for $p = 2$.

Theorem 5.8: Consider the case of $p = 2$, the values of the weights W_{cjl} 's that minimize $V(\widehat{Y}_{MDcj})$ is given by the expressions

$$W_{cj1} = \frac{(V_{cj2} - V_{cj12})}{(V_{cj1} - 2V_{cj12} + V_{cj2})} \quad (5.53)$$

and

$$W_{cj2} = \frac{(V_{cj1} - V_{cj12})}{(V_{cj1} - 2V_{cj12} + V_{cj2})} \quad (5.54)$$

Proof. In this proof, the problem is finding the expressions for the values of W_{cjl} 's that minimize $V(\widehat{Y}_{MDcj})$ subject to the linear condition $\sum_{l=1}^2 W_{cjl} = 1$. Considering a particular response group $j = 1$, (say), so that the focus is to

minimize $V(\widehat{Y}_{MDc1})$ subject to the condition that $W_{c11} + W_{c12} = 1$. To do this, we conditionally minimize the function

$$\Phi = V(\widehat{Y}_{MDc1}) + \lambda(1 - W_{c11} - W_{c12}) \quad (5.55)$$

where λ is the Lagrange's Multiplier.

From equation (5.43), $V(\widehat{Y}_{MDc1})$ is expressed as

$$V(\widehat{Y}_{MDc1}) = W_{c11}^2 V_{c11} + 2W_{c11}W_{c12}V_{c112} + W_{c12}^2 V_{c12} \quad (5.56)$$

which, when replaced in equation (5.55) gives

$$\Phi = W_{c11}^2 V_{c11} + 2W_{c11}W_{c12}V_{c112} + W_{c12}^2 V_{c12} + \lambda(1 - W_{c11} - W_{c12}) \quad (5.57)$$

To minimize Φ , equation (5.57) is differentiated partially with respect to W_{c11} and W_{c12} separately and equated to 0 as shown below

$$\frac{\partial \Phi}{\partial W_{c11}} = 2W_{c11}V_{c11} + 2W_{c12}V_{c112} - \lambda \quad (5.58)$$

and

$$\frac{\partial \Phi}{\partial W_{c12}} = 2W_{c12}V_{c12} + 2W_{c11}V_{c112} - \lambda \quad (5.59)$$

Equating both equations (5.58) and (5.59) to zero gives

$$W_{c11}(V_{c11} - V_{c112}) = W_{c12}(V_{c12} - V_{c112}) \quad (5.60)$$

But $W_{c12} = 1 - W_{c11}$ so that equation (5.60) can be expressed as

$$W_{c11}(V_{c11} - V_{c112}) = (1 - W_{c11})(V_{c12} - V_{c112})$$

which implies that

$$W_{c11}\{(V_{c11} - V_{c112}) + (V_{c12} - V_{c112})\} = (V_{c12} - V_{c112}) \quad (5.61)$$

From equation (5.61), we have

$$\begin{aligned} W_{c11} &= \frac{(V_{c12} - V_{c112})}{\{(V_{c11} - V_{c112}) + (V_{c12} - V_{c112})\}} \\ &= \frac{(V_{c12} - V_{c112})}{(V_{c11} - 2V_{c112} + V_{c12})} \end{aligned} \quad (5.62)$$

On the other hand, the value of W_{c12} that minimizes Φ is obtained using the linear condition $W_{c12} = 1 - W_{c11}$ and is thus given by

$$W_{c12} = \frac{(V_{c11} - V_{c112})}{(V_{c11} - 2V_{c112} + V_{c12})} \quad (5.63)$$

Now, using a similar procedure for $j = 2$, we obtain

$$W_{c21} = \frac{(V_{c22} - V_{c212})}{(V_{c21} - 2V_{c212} + V_{c22})} \quad (5.64)$$

and

$$W_{c22} = \frac{(V_{c21} - V_{c212})}{(V_{c21} - 2V_{c212} + V_{c22})} \quad (5.65)$$

Therefore, from equations (5.62), (5.63), (5.64) and (5.65), it can be seen that for $p = 2$,

$$W_{cj1} = \frac{(V_{cj2} - V_{cj12})}{(V_{cj1} - 2V_{cj12} + V_{cj2})} \text{ and } W_{cj2} = \frac{(V_{cj1} - V_{cj12})}{(V_{cj1} - 2V_{cj12} + V_{cj2})}$$

Hence the proof. □

Corollary 5.2: From Theorem 5.8, the minimum variance of \widehat{Y}_{MDcj} at the optimal values of W_{c11} and W_{c12} , denoted by $V_{min}(\widehat{Y}_{MDcj})$, is given by

$$V_{min}(\widehat{Y}_{MDcj}) = \frac{V_{cj1}V_{cj2} - V_{cj12}^2}{V_{cj1} + V_{cj2} - 2V_{cj12}} \quad (5.66)$$

Proof. Using the expressions for W_{cj1} and W_{cj2} given in equations (5.53) and (5.54), the expression for the minimum variance of \widehat{Y}_{MDcj} is given by

$$V_{min}(\widehat{Y}_{MDcj}) = \frac{V_{cj1}V_{cj2} - V_{cj12}^2}{V_{cj1} + V_{cj2} - 2V_{cj12}}$$

Now, from equation (5.43),

$$V(\widehat{Y}_{MDcj}) = W_{cj1}^2 V_{cj1} + 2W_{cj1}W_{cj2}V_{cj12} + W_{cj2}^2 V_{cj2} \quad (5.67)$$

Substituting the expressions for W_{cj1} and W_{cj2} given as

$$W_{cj1} = \frac{(V_{cj2} - V_{cj12})}{(V_{cj1} - 2V_{cj12} + V_{cj2})} \text{ and } W_{cj2} = \frac{(V_{cj1} - V_{cj12})}{(V_{cj1} - 2V_{cj12} + V_{cj2})}$$

in equation (5.67) and simplifying gives

$$V_{min}(\widehat{Y}_{MDcj}) = \frac{V_{cj1}V_{cj2} - V_{cj12}^2}{V_{cj1} + V_{cj2} - 2V_{cj12}}$$

Hence the proof. \square

Corollary 5.3: An unbiased estimator for $V_{min}(\widehat{Y}_{MDcj})$ for $p = 2$ is given by

$$V_{min}(\widehat{Y}_{MDcj}) = \frac{v_{cj1}v_{cj2} - v_{cj12}^2}{v_{cj1} + v_{cj2} - 2v_{cj12}} \quad (5.68)$$

where v_{cj1} , v_{cj2} and v_{cj12} are unbiased estimators for V_{cj1} , V_{cj2} and V_{cj12} respectively.

Corollary 5.4: If $V_{cjl'l'} = \text{Covariance}(\widehat{Y}_{Dcjl}, \widehat{Y}_{Dcjl'}) = 0$ for any $l \neq l'$, variance of \widehat{Y}_{MD} reduces to

$$V(\widehat{Y}_{MD}) = \sum_{c=1}^k \sum_{j=1}^2 \sum_{l=1}^p W_{cjl}^2 V_{cjl} \quad (5.69)$$

Corollary 5.5: If $V_{cjl'l'} = 0$ for any $l \neq l'$ and $p = 2$, then

$$W_{cj1} = \frac{V_{cj2}}{(V_{cj1} + V_{cj2})} \text{ and } W_{cj2} = \frac{V_{cj1}}{(V_{cj1} + V_{cj2})} \quad (5.70)$$

and the corresponding $V_{min}(\widehat{Y}_{MDcj})$ is given by

$$V_{min}(\widehat{Y}_{MDcj}) = \frac{V_{cj1}V_{cj2}}{V_{cj1} + V_{cj2}} \quad (5.71)$$

Now, for the general multivariate form, \widehat{Y}_{MD} , we conditionally minimize the function

$$\Theta = \sum_{c=1}^k \sum_{j=1}^2 \left(\sum_{l=1}^p W_{cjl}^2 V_{cjl} + 2 \sum_{l=1}^{p-1} W_{cjl} W_{cjl'} V_{cjl'l'} + \lambda \left(1 - \sum_{l=1}^p W_{cjl} \right) \right) \quad (5.72)$$

where λ is the Lagrange's Multiplier.

Equation (5.72) is minimized by differentiating Θ partially with respect to W_{cjl} and equating to zero to get

$$\frac{\partial \Theta}{\partial W_{cjl}} = 2W_{cjl}V_{cjl} + 2 \sum_{l=1}^{p-1} W_{cjl'}V_{cjl'l'} - \lambda \text{ for } l = 1, 2, \dots, p \quad (5.73)$$

In particular,

$$\begin{cases} \frac{\partial \Theta}{\partial W_{c j 1}} = 2W_{c j 1}V_{c j 1} + 2W_{c j 2}V_{c j 1 2} + 2W_{c j 3}V_{c j 1 3} + \cdots + 2W_{c j p}V_{c j 1 p} - \lambda \\ \frac{\partial \Theta}{\partial W_{c j 2}} = 2W_{c j 1}V_{c j 2 1} + 2W_{c j 2}V_{c j 2} + 2W_{c j 3}V_{c j 2 3} + \cdots + 2W_{c j p}V_{c j 2 p} - \lambda \\ \frac{\partial \Theta}{\partial W_{c j 3}} = 2W_{c j 1}V_{c j 3 1} + 2W_{c j 2}V_{c j 3 2} + 2W_{c j 3}V_{c j 3} + \cdots + 2W_{c j p}V_{c j 3 p} - \lambda \\ \vdots \\ \frac{\partial \Theta}{\partial W_{c j p}} = 2W_{c j 1}V_{c j p 1} + 2W_{c j 2}V_{c j p 2} + 2W_{c j 3}V_{c j p 3} + \cdots + 2W_{c j p}V_{c j p} - \lambda \end{cases} \quad (5.74)$$

Equating the set of equations (5.74) to zero gives,

$$\begin{bmatrix} 2V_{c j 1} & 2V_{c j 1 2} & 2V_{c j 1 3} & \cdots & 2V_{c j 1 p} \\ 2V_{c j 2 1} & 2V_{c j 2} & 2V_{c j 2 3} & \cdots & 2V_{c j 2 p} \\ 2V_{c j 3 1} & 2V_{c j 3 2} & 2V_{c j 3} & \cdots & 2V_{c j 3 p} \\ \vdots & & & & \vdots \\ 2V_{c j p 1} & 2V_{c j p 2} & 2V_{c j p 3} & \cdots & 2V_{c j p} \end{bmatrix} \begin{bmatrix} W_{c j 1} \\ W_{c j 2} \\ W_{c j 3} \\ \vdots \\ W_{c j p} \end{bmatrix} = \begin{bmatrix} \lambda \\ \lambda \\ \lambda \\ \vdots \\ \lambda \end{bmatrix}$$

Or equivalently,

$$\begin{bmatrix} V_{c j 1} & V_{c j 1 2} & V_{c j 1 3} & \cdots & V_{c j 1 p} \\ V_{c j 2 1} & V_{c j 2} & V_{c j 2 3} & \cdots & V_{c j 2 p} \\ V_{c j 3 1} & V_{c j 3 2} & V_{c j 3} & \cdots & V_{c j 3 p} \\ \vdots & & & & \vdots \\ V_{c j p 1} & V_{c j p 2} & V_{c j p 3} & \cdots & V_{c j p} \end{bmatrix} \begin{bmatrix} W_{c j 1} \\ W_{c j 2} \\ W_{c j 3} \\ \vdots \\ W_{c j p} \end{bmatrix} = \begin{bmatrix} \frac{\lambda}{2} \\ \frac{\lambda}{2} \\ \frac{\lambda}{2} \\ \vdots \\ \frac{\lambda}{2} \end{bmatrix} \quad (5.75)$$

which is a consistent linear system of p linear equations. A consistent linear system is a linear system with unique solutions.

The augmented matrix of the linear system given in equation (5.75) is given as

$$\left[\begin{array}{cccc|c} V_{cj1} & V_{cj12} & V_{cj13} & \cdots & V_{cj1p} & q \\ V_{cj21} & V_{cj2} & V_{cj23} & \cdots & V_{cj2p} & q \\ V_{cj31} & V_{cj32} & V_{cj3} & \cdots & V_{cj3p} & q \\ \vdots & & & & \vdots & \vdots \\ V_{cjp2} & V_{cjp2} & V_{cjp3} & \cdots & V_{cjp} & q \end{array} \right] \quad (5.76)$$

where $q = \frac{\lambda}{2}$, while the coefficient matrix is given as

$$\begin{bmatrix} V_{cj1} & V_{cj12} & V_{cj13} & \cdots & V_{cj1p} \\ V_{cj21} & V_{cj2} & V_{cj23} & \cdots & V_{cj2p} \\ V_{cj31} & V_{cj32} & V_{cj3} & \cdots & V_{cj3p} \\ \vdots & & & & \vdots \\ V_{cjp2} & V_{cjp2} & V_{cjp3} & \cdots & V_{cjp} \end{bmatrix}$$

Now, equation (5.75) can be expressed as

$$\underline{\mathbf{V}} \underline{\mathbf{W}} = \underline{\lambda} \quad (5.77)$$

so that using row operations, values of the weights can be obtained from

$$\underline{\mathbf{I}} \underline{\mathbf{W}} = \underline{\mathbf{V}}^{-1} \underline{\lambda}, \quad (5.78)$$

where $\underline{\mathbf{I}}$ is an identity matrix.

For known values of V_{cjl} and $V_{cjl'}$ (where $l = 1, 2, 3, \dots, p$ and $l \neq l'$) row operations can involve any or all of the following

- i) Adding a multiple of one row to another row(s)
- ii) Interchanging two rows
- iii) Multiplying all entries of one row by a non-zero constant

Using these row operations to reduce equation (5.79) to

$$\left[\begin{array}{cccc|c} 1 & 0 & 0 & \cdots & 0 & d_1 \\ 0 & 1 & 0 & \cdots & 0 & d_2 \\ 0 & 0 & 1 & \cdots & 0 & d_3 \\ \vdots & & & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & d_p \end{array} \right] \quad (5.79)$$

gives the solution of the linear system (5.75) as $W_{cjl} = d_l$. That is, $W_{cj1} = d_1, W_{cj2} = d_2, \dots, W_{cjp} = d_p$.

5.5 Simulation Study

In this section, performances of the estimators \hat{Y}_D , \hat{Y}_{DC} , \hat{Y}_{DR} and \hat{Y}_{MD} are verified using simulated data. A hypothetical population of 300 units consisting of three strata ($c = 1, 2, 3$), where the stratum sizes are randomly generated such that the sum equals to 300, is considered. Krejcie-Morgan-Sample-Size-Table is used to get overall sample size as 170 and the stratum sample sizes are allocated using proportional allocation technique. For univariate estimation, normally distributed random vectors for the auxiliary variable X and for the response variable Y are generated and a linear model of Y on X is fit in each stratum using the linear regression model in R. For multivariate estimation, another auxiliary variable is added so that we have a two-component auxiliary random vector. A random number generator is then used to identify sample units in each response group in all the three strata. That is, sampling is done index-wise such that if i^{th} index is selected then the sample element will be the i^{th} pair (X_i, Y_i) . For the non-response, a non-response rate of 20% is assumed. Using this procedure, population and sample sizes are obtained as shown in Table 5.1.

Table 5.1: Stratum Population and Sample Sizes

Stratum Id.	Response Group	Pop. Size, N_{cj}	Sample Size, n_{cj}	Subsample, m_c
Stratum 1	j=1	133	75	-
	j=2	34	19	16
Stratum 2	j=1	75	42	-
	j=2	19	11	10
Stratum 3	j=1	31	18	-
	j=2	8	5	4

From Table 5.1, it is observed that in each stratum, a sub-sample of size m_c is obtained from the n_{c2} non-responding units and an assumption that the sub-sampled units will fully respond is made.

Next is to show how the random data set was generated from normal population with different parameters as shown in Table 5.2

From Table 5.2, both the response variable and the auxiliary variables are from normal population with different parameters in each stratum. The weights column gives randomly generated values from uniform distribution for each stratum using R software. Values of the weights in multivariate

Table 5.2: Random Data Parameters

Stratum	Study Variable, Y	Auxiliary Random Vector, \underline{X}		Weights	
		p=1, X_1	p=2, X_2	W_{cj1}	W_{cj2}
Stratum 1	rnorm(167, 35, 12)	rnorm(167, 45, 14.5)	rnorm(167, 43, 18)	0.76	0.24
Stratum 2	rnorm(94, 35, 11)	rnorm(94, 40, 12)	rnorm(94, 42, 14)	0.67	0.33
Stratum 3	rnorm(39, 48, 14)	rnorm(39, 32, 16)	rnorm(39, 45, 14.5)	0.50	0.50

ratio estimation can be obtained using randomly generated values such that the sum of the weights is 1 (one) or utilizing sample information using expression given in equation (5.68). The weights shall be used in multivariate estimation of population total for a two-component auxiliary random vector.

We shall begin with univariate estimation of finite population total using one auxiliary variable X_1 . In univariate estimation, efficiency of the estimators \hat{Y}_D , \hat{Y}_{DC} and \hat{Y}_{DR} is compared with the usual ratio estimator under SRSWOR and under stratified random sampling. The following estimators are, thus, considered:

Ratio estimator under SRSWOR, $t_1 = \sum_{j=1}^2 r_j X_{Tj}$, where $r_j = \frac{\bar{y}_j}{\bar{x}_j}$

Ratio estimator under stratified sampling, $t_2 = \sum_{j=1}^2 \sum_{c=1}^k r_{cj} X_{Tcj}$, where $r_{cj} = \frac{\bar{y}_{cj}}{\bar{x}_{cj}}$

Unbiased ratio estimator under SRSWOR, $\hat{Y}_{D(SRS)} = \sum_{j=1}^2 \left[\bar{r}_j X_{Tj} + \frac{N_j-1}{n_j-1} (y_{tj} - \bar{r}_j x_{tj}) \right]$,

where $\bar{r}_j = \frac{\bar{y}_j}{\bar{x}_j}$

Unbiased separate ratio estimator, $\hat{Y}_D = \sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_{cj} X_{Tcj} + \frac{N_{cj}-1}{n_{cj}-1} (y_{tcj} - \bar{r}_{cj} x_{tcj}) \right]$

Unbiased combined ratio estimator, $\hat{Y}_{DC} = \sum_{c=1}^k \sum_{j=1}^2 \left[\bar{r}_j^* X_{Tcj} + \frac{N_{cj}-1}{n_{cj}-1} \{y_{tcj} - \bar{r}_j^* x_{tcj}\} \right]$,

where $\bar{r}_j^* = \frac{\sum_{c=1}^k n_{cj} \bar{y}_{cj}}{\sum_{c=1}^k n_{cj} \bar{x}_{cj}}$

Unbiased regression estimator, $\widehat{Y}_{DR} = \sum_{j=1}^2 \sum_{c=1}^k N_{cj} \left[\widehat{Y}_{cj}^* + \beta_{0cj} (\overline{X}_{cj} - \overline{x}_{cj}) \right]$

with a corresponding variances:

$$Var(t_1) = \sum_{j=1}^2 \frac{N_j(N_j - n_j)}{n_j} \left[S_{yj}^2 + R_j^2 S_{xj}^2 - 2R_j S_{xyj} \right]$$

$$Var(t_2) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + R_{cj}^2 S_{xcj}^2 - 2R_{cj} S_{xycj} \right]$$

$$Var(\widehat{Y}_{D(SRS)}) = \sum_{j=1}^2 \frac{N_j(N_j - n_j)}{n_j} \left[S_{yj}^2 + \overline{R}_j^2 S_{xj}^2 - 2\overline{R}_j S_{xyj} \right. \\ \left. + \frac{1}{n_j - 1} \{ S_{rj}^2 S_{xj}^2 + S_{rxj} \} \right]$$

$$Var(\widehat{Y}_D) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \overline{R}_j^2 S_{xcj}^2 - 2\overline{R}_j S_{xycj} \right. \\ \left. + \frac{1}{n_{cj} - 1} \{ S_{rj}^2 S_{xcj}^2 + S_{rxcj} \} \right]$$

$$Var(\widehat{Y}_{DC}) = \sum_{c=1}^k \sum_{j=1}^2 \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{ycj}^2 + \overline{R}_j^{*2} S_{xcj}^2 - 2\overline{R}_j^* S_{xycj} \right. \\ \left. + \frac{1}{n_{cj} - 1} \{ S_{r^*j}^2 S_{xcj}^2 + S_{r^*xcj} \} \right]$$

$$Var(\widehat{Y}_{DR}) = \sum_{j=1}^2 \sum_{c=1}^k \frac{N_{cj}(N_{cj} - n_{cj})}{n_{cj}} \left[S_{Ycj}^2 + \beta_{0cj}^2 S_{Xcj}^2 - 2\beta_{0cj} S_{XYcj} \right]$$

Performances of the estimators are examined at four different levels. The first level involves comparing the adopted sub-sampling method with another non-response correction method, which is partial deletion method. The second level of comparison involves comparing performance of \widehat{Y}_D under SRSWOR ($\widehat{Y}_{D(SRS)}$) and under stratified random sampling. The third level of comparison entails examining the performance of different forms of \widehat{Y}_D constructed under stratified random sampling. Here, efficiency of \widehat{Y}_D , \widehat{Y}_{DC} and \widehat{Y}_{DR} is examined relative to $\widehat{Y}_{D(SRS)}$. The last level of comparison shall involve comparing the performance of the multivariate unbiased ratio estimator \widehat{Y}_{MR} relative to the multivariate estimators by Olkin (1958) and Ngesa et al. (2012). For univariate estimation, only the first three levels of

comparisons are considered.

Now for the univariate, the summary statistics for the response and the auxiliary variables were obtained as shown in Table 5.3 and Table 5.4.

Table 5.3: Summary Statistics for SRSWOR

Statistics	Response Group		Total
	j=1	j=2	
X_{Tj}	10694.11	2665.41	13359
\bar{r}_j	0.7399	0.7169	
S_{xj}^2	207.59	165.02	372.61
S_{yj}^2	166.46	147.09	313.55
S_{rj}^2	0.282	0.103	0.385
S_{xyj}	13.10	10.44	23.54
S_{rxj}	-4.79	-2.46	-7.25

Table 5.3 shows summary statistics for SRSWOR for computation of t_1 and $\hat{Y}_{D(SRS)}$. That is, the table gives the summary statistics if the three strata are collapsed. It can be noted from Table 5.3 that for the population totals of the auxiliary variable were found to be 10694.11 and 2665.41 for the responding and non-responding groups respectively. The table also gives sample estimates for S_{xj}^2 , S_{yj}^2 , S_{rj}^2 , S_{xyj} and S_{rxj} . All the values in Table 5.3 are used to compute t_1 and $Y_{D(SRS)}$. For stratified population, summary statistics are as shown in Table 5.4.

Table 5.4: Summary Statistics

Statistics.	Stratum 1		Stratum 2		Stratum 3		Total
	j=1	j=2	j=1	j=2	j=1	j=2	
y_{tcj}	2714.32	480.25	1474.99	381.09	514.80	116.69	5682.14
x_{tcj1}	3522.63	729.60	1859.57	403.89	907.67	234.98	7658.34
X_{Tcj1}	6193.24	1529.88	3019.30	732.60	1481.57	402.93	13359.52
\bar{y}_{cj}	36.19	30.02	35.12	38.11	28.60	29.17	
\bar{x}_{cj1}	46.97	45.60	44.28	40.39	50.43	58.75	
\bar{r}_{cj1}	0.7000	0.6915	0.7599	0.8322	0.6399	0.5520	
r_{cj1}	0.7705	0.6582	0.7932	0.9436	0.5672	0.4966	
\bar{r}_{j1}^*	$j = 1, r_{11} = 0.7479$			$j = 2, r_{12} = 0.7147$			
s_{ycj}^2	139.30	166.99	132.23	112.17	337.28	97.90	985.8797
s_{xcj1}^2	244.02	157.49	171.40	98.00	131.30	192.75	994.9591
s_{rcj1}^2	0.1664	0.0840	0.5177	0.0902	0.1670	0.0590	1.0843
s_{xycj1}	37.50	52.367	-16.361	28.129	4.146	-89.56	16.221
s_{rxcj1}	-4.26	-1.619	-6.584	-1.628	-1.837	-3.09	-19.018
ρ_{xycj}	0.2034	0.3229	-0.1087	0.2683	0.0197	-0.652	
β_{cj}	0.1537	0.3325	-0.095	0.287	0.0316	-0.465	

Table 5.4 gives the estimates of population totals, ratio means, variances, covariances, correlation coefficients and regression coefficients for all the strata in each response group. Though the correlation coefficients in each stratum ρ_{xycj} are not zero, the values, however, indicate a weak relationship between X and Y in each stratum. A look at the values of ρ_{xycj} and β_{cj} shows that the paired values have the same sign and this supports existence of linear relationship between X and Y . For example, in stratum 1 at $j = 1$, both values are positive, while in stratum 2 at $j = 1$, both values are negative. It is worth noting that the subscript 1 (for example s_{xcj1}^2 or \bar{x}_{cj1}) is used to imply statistics computed with respect to the first component of the auxiliary random vector. All values in Table 5.4 are used to examine performances of t_2 , $\hat{Y}_{D(SRS)}$, \hat{Y}_D , \hat{Y}_{DC} and \hat{Y}_{DR} whose statistics are given in Table 5.5.

Table 5.5: Results of Univariate Estimation

	Stratum 1		Stratum 2		Stratum 3		Total
	j=1	j=2	j=1	j=2	j=1	j=2	
t_1	$j = 1, t_{11} = 7997.06$		$j = 2, t_{12} = 1904.70$				9901.75
$\hat{Y}_{D(SRS)}$	$j = 1, \hat{Y}_{D1(SRS)} = 8000.82,$		$j = 2, Y_{D2(SRS)} = 1904.32$				9905.14
t_2	4772.10	1007.00	2394.90	691.24	840.30	200.10	9905.66
\hat{Y}_D	4778.50	1004.50	2406.10	699.62	831.56	192.04	9912.32
\hat{Y}_{DC}	4774.50	1003.178	2409.87	708.84	818.48	169.15	9883.60
\hat{Y}_{DR}	4770.27	997.70	2434.86	689.63	828.98	223.19	9944.62
MSE(t_1)	$j = 1, \text{MSE}(t_{11}) = 48414.51$		$j = 2, \text{MSE}(t_{12}) = 13642.42$				62056.94
MSE($\hat{Y}_{D(SRS)}$)	$\text{MSE}(\hat{Y}_{D1(SRS)}) = 48077.27$		$\text{MSE}(\hat{Y}_{D2(SRS)}) = 13705.25$				61782.51
MSE(t_2)	23291.31	6362.19	15683.62	2504.44	8392.70	1885.05	58119.32
MSE(\hat{Y}_D)	21276.40	6527.39	15208.00	2291.84	8662.73	2066.14	56032.53
MSE(\hat{Y}_{DC})	22671.47	6607.19	15015.77	2092.86	9086.66	2598.21	58072.16
MSE(\hat{Y}_{DR})	13734.70	5721.54	7700.27	1780.02	7548.42	450.27	36935.20
PRE(t_1)							100.00%
PRE($\hat{Y}_{D(SRS)}$)							100.44%
PRE(t_2)							106.78%
PRE(\hat{Y}_D)							110.26%
PRE(\hat{Y}_{DC})							106.39%
PRE(\hat{Y}_{DR})							167.62%

From Table 5.5, it is observed that the estimates of population total using different estimators do not vary significantly from one another with the least estimate being $\hat{Y}_{DC} = 9883.60$ and the highest being $\hat{Y}_{DR} = 9944.62$. From the simulated data, the true population total is 10209.35. Now, comparing the estimates from different estimators and the true population total, it can be noted that \hat{Y}_{DR} has the least absolute deviation from the true mean of 264.73 while \hat{Y}_{DC} has the highest absolute deviation from the true mean of 325.75. Therefore, based on bias alone, one can prefer \hat{Y}_{DR} , the regression estimator, to other ratio-type estimators. This conclusion is, however, dependent on the MSE's of the estimators.

Though in our theoretical proofs, the estimators \hat{Y}_D , \hat{Y}_{DC} and \hat{Y}_{DR} were all shown to be unbiased for finite population total, the population total es-

timates, however, indicate otherwise. This can be attributed to the weak correlation between X and Y as presented in Table 5.4. As previously mentioned that the use of auxiliary variable in estimation of population parameters is based on the existence of a perfect linear relationship between X and Y such that the regression line of Y on X passes through the origin. That is, efficiency of any ratio-type estimator is improved if the correlation coefficients ρ_{xy} 's is close to one. This condition is not met based on the coefficient values given in Table 5.4.

To examine the appropriateness of the adopted sub-sampling method in this study, PRE's of t_1 and $\hat{Y}_{D(SRS)}$ are compared. By definition, the percent efficiency of $\hat{Y}_{D(SRS)}$ relative to t_1 is obtained as follows

$$\begin{aligned} \text{PRE}(\hat{Y}_{D(SRS)}) &= \frac{\text{MSE}(t_1)}{\text{MSE}(\hat{Y}_{D(SRS)})} \times 100 \\ &= \frac{62056.94}{61782.51} \times 100 \\ &= 100.44\% \end{aligned}$$

The value 100.44% implies that based on the sampled data, the gain in precision if sub-sampling method is used compared to partial deletion would be 0.44%. Though this value is small, there is, however, gain in precision since 100.44% > 100.00%.

For comparison of the sampling scheme, performances of \hat{Y}_D under SR-SWOR and under stratified random sampling are compared. That is, for this procedure, $\text{MSE}(\hat{Y}_{D(SRS)})$ and $\text{MSE}(\hat{Y}_D)$ are compared. From Table 5.5, $\text{MSE}(\hat{Y}_{D(SRS)}) = 61782.51$ and $\text{MSE}(\hat{Y}_D) = 56032.56$, which implies that \hat{Y}_D is more efficient than $\hat{Y}_{D(SRS)}$ and the gain in precision is

$$\left[\frac{61782.51}{56032.53} \times 100 \right] - 100 = 10.26\%.$$

The value 10.26% implies that stratification improves precision of \hat{Y}_D by 10.26%. A similar conclusion can be made by comparing t_1 and t_2 , with the respective MSE's being 62056.94 and 58119.32. In this case, the gain in precision due to stratification is

$$\left[\frac{62056.94}{58119.32} \times 100 \right] - 100 = 6.8\%.$$

This observation does not contradict a known knowledge that stratification improves efficiency estimators (Cochran, 1977).

In each stratum, comparison of MSEs of t_2 , \hat{Y}_D and \hat{Y}_{DC} in Table 5.5 with respect to the values of r_{cj1} , \bar{r}_{j1}^* and \bar{r}_{cj1} in Table 5.4 reveals some underlying trend. Results in these two tables show that in each stratum and each response group, the smaller the value of any of the ratios, the smaller the corresponding MSE of the population estimate. For instance, for stratum 1, response group 1, $r_{111} = 0.7705$, $\bar{r}_{11}^* = 0.7479$ and $\bar{r}_{111} = 0.7000$ and the corresponding MSEs are $MSE(t_2) = 23291.31$, $MSE(\hat{Y}_{DC}) = 22647.59$ and $MSE(\hat{Y}_D) = 21276.40$. This trend is repeated in all the strata and in all response groups.

A similar conclusion can be made when $MSE(\hat{Y}_{DC})$ and $MSE(\hat{Y}_D)$ are compared with respect to \bar{r}_j^* and \bar{r}_{cj} . In Section 4.5, it was proved that efficiency of either \hat{Y}_D or \hat{Y}_{DC} depends on the values of \bar{R}_j^* and \bar{R}_{cj} . The simulated data supports this proof by showing that $MSE(\hat{Y}_{DC})$ and $MSE(\hat{Y}_D)$ depends on the values of \bar{r}_j^* and \bar{r}_{cj} . For the unbiased regression estimator, \hat{Y}_{DR} produces efficient estimator in all the strata and response groups.

For different forms of \hat{Y}_D , MSE's of \hat{Y}_D , \hat{Y}_{DC} and \hat{Y}_{DR} are compared using $\hat{Y}_{D(SRS)}$ as the base estimator. The respective PRE's for \hat{Y}_D , \hat{Y}_{DC} and \hat{Y}_{DR} are 110.26%, 106.39% and 167.62%. These values imply that the regression estimator \hat{Y}_{DR} performs better than \hat{Y}_D and \hat{Y}_{DC} , with a gain in precision of 67.62%. This observation supports the proof in Section 5.3, where we showed that \hat{Y}_{DR} performs better than other ratio-type estimators in both SRSWOR and stratified random sampling. Now, relating the absolute deviations of estimates given by \hat{Y}_D , \hat{Y}_{DC} and \hat{Y}_{DR} and the corresponding PRE's, it can be observed that \hat{Y}_{DR} has the smallest absolute difference but a higher PRE, while \hat{Y}_{DC} has the largest absolute difference but the least PRE. This relation implies that the regression form of the estimator \hat{Y}_D performs better than other unbiased ratio-type estimators in estimating finite population total in stratified random sampling under non-response.

From the results of the univariate estimation, it can be concluded that subsampling method suggested by Hansen and Hurwitz (1946) produces more efficient estimators than partial deletion method of correcting non-response.

It has also been observed that stratification improves efficiency of estimators compared to SRSWOR. Further, the results have indicated that performance of separate ratio estimator over combined ratio estimator depends on the absolute difference between the combined ratio and the individual ratios in each stratum. In addition, univariate results have confirmed that regression estimation produces more efficient estimators compared to SRSWOR and stratified random sampling. All these observations are consistent with the known literature knowledge on ratio estimation.

For the fourth level of comparison, performance of the multivariate estimator \hat{Y}_{MD} for finite population total is examined using a two-component auxiliary random vector. From equation (5.33), the overall multivariate estimator of finite population total is a linear function of the population total estimator obtained using each component of the auxiliary random vector pre-multiplied by the weights W_{cjl} for $l = 1, 2, \dots, p$. Values in Table 5.4 and Table 5.5 give summary statistics with respect to the first component (X_1) of the auxiliary random vector (X). For the second component of the auxiliary random vector, summary statistics is given in Table 5.6.

Table 5.6: Summary Statistics for X_2

	Stratum 1		Stratum 2		Stratum 3		Total
	j=1	j=2	j=1	j=2	j=1	j=2	
x_{tcj2}	3066.19	669.83	1754.97	478.26	811.11	183.61	6963.97
\bar{x}_{cj2}	40.88	41.86	41.79	47.83	45.06	45.90	
X_{Tcj2}	5464.19	1360.60	3164.01	808.27	1469.63	352.06	12618.76
\bar{r}_{cj2}	0.8583	0.8065	0.7909	0.7826	0.5883	0.6052	
r_{cj2}	0.8853	0.7172	0.8404	0.7968	0.6346	0.6355	
r_{j2}	$j = 1, r_{12} = 0.8206$			$j = 2, r_{22} = 0.7344$			
$s_{x_{cj2}}^2$	261.97	328.95	211.95	143.33	139.74	399.53	1485.47
$s_{r_{cj2}}^2$	0.781	0.556	0.253	0.0073	0.193	0.168	2.034
$s_{x_{ycj2}}$	-7.54	100.38	12.83	23.63	16.79	15.01	161.1
$s_{r_{xcj2}}$	-10.79	-8.47	4.72	-1.90	-2.09	-6.09	-34.06

From Table 5.6, the subscript 2 is used to indicate that the values in the table are computed with respect to X_2 . Therefore, to obtain the multivariate estimate of finite population total using a two-component auxiliary random

vector, the values in Table 5.4, Table 5.6 and the weights in Table 5.3 are used. While computing $\text{Var}(\widehat{Y}_{MD})$, an assumption that $\text{Cov}(Y_{cjl}, Y_{cjl'}) = 0$, for all $l \neq l'$ is made, so that

$$V(\widehat{Y}_{MD}) = \sum_{c=1}^k \sum_{j=1}^2 \left[W_{cj1}^2 V_{cj1} + W_{cj2}^2 V_{cj2} \right].$$

A similar procedure is used to obtain the estimate of finite population total using a multivariate form of \widehat{Y}_{DC} , which shall be denoted as \widehat{Y}_{MDC} . Therefore, performances of \widehat{Y}_{MD} and \widehat{Y}_{MDC} are compared against the multivariate ratio estimator by Olkin (1958) and by Ngesa et al. (2012), denoted as t_O and t_N respectively. For Olkin's estimator (1958), the weights are obtained from the averages of the randomly generated weights such that for $j = 1$, $W_1 = 0.64$ and for $j = 2$, $W_1 = 0.36$. The four multivariate estimators are, thus, expressed as:

$$t_O = \sum_{l=1}^2 W_l \frac{\bar{y}}{\bar{x}_l} X_l = \sum_{l=1}^2 W_l r_l X_l$$

$$t_N = \sum_{c=1}^k \widehat{Y}_{MRc}$$

$$\widehat{Y}_{MD} = \sum_{c=1}^k \sum_{j=1}^2 \sum_{l=1}^2 W_{cjl} [\bar{r}_{cjl} X_{Tcjl} + \frac{N_{cj}-1}{n_{cj}-1} (y_{tcj} - \bar{r}_{cjl} x_{tcjl})]$$

$$\widehat{Y}_{MDC} = \sum_{c=1}^k \sum_{j=1}^2 \sum_{l=1}^2 W_{cjl} [\bar{r}_{cjl}^* X_{Tcjl} + \frac{N_{cj}-1}{n_{cj}-1} (y_{tcj} - \bar{r}_{cjl}^* x_{tcjl})]$$

Using these multivariate estimators and the randomly generated weights given in Table 5.2, summary statistics are obtained as given in Table 5.7

Table 5.7: Results for Multivariate Estimation

	Stratum 1		Stratum 2		Stratum 3		Total
	j=1	j=2	j=1	j=2	j=1	j=2	
t_O	$j = 1, t_{O1} = 6337.98$		$j = 2, t_{O2} = 3649.55$				9987.53
t_N	4392.12	1385.43	2067.71	1090.07	520.2	578.25	10033.77
\hat{Y}_{MD}	4395.08	1392.65	2080.83	1090.84	511.80	578.73	10049.72
\hat{Y}_{MDC}	4390.48	1394.64	2089.45	1093.17	493.20	578.04	10038.97
MSE(t_O)	$j = 1, \text{MSE}(t_{O1}) = 25418.52$		$j = 2, \text{MSE}(t_{O2}) = 9368.08$				34786.61
MSE(t_N)	16911.3	2543.93	8164.61	1979.19	2569.44	2564.02	34732.49
MSE(\hat{Y}_{MD})	16059.47	2569.46	7855.69	1882.06	2682.22	2548.93	33597.88
MSE(\hat{Y}_{MDC})	16911.3	2419.24	7677.25	1926.99	2925.09	2891.85	34751.72
PRE(t_O)							100.00%
PRE(t_N)							100.16%
PRE(\hat{Y}_{MD})							103.54%
PRE(\hat{Y}_{MDC})							100.10%

Table 5.7 gives population total estimates as given by the aforementioned multivariate estimators. Corresponding variances and the percent relative efficiency with respect to the Olkin's estimator (1958) are also given in the table. A look at the population total estimates in Table 5.7 reveals that there is no big difference in the multivariate estimates except for Olkin's estimator that has the least value. A look at absolute differences between the estimates and the actual population total (which is 10319.35) in both univariate (Table 5.5) and multivariate results shows that multivariate estimators generally have less absolute deviations. But this observation is, nevertheless, based on the values of the weights.

It is worth noting that Olkin's multivariate estimator is calculated under SR-SWOR using partial deletion to take care of the non-response. To check for appropriateness of sub-sampling method suggested by Hansen and Hurwitz (1946) over partial deletion, MSE's of t_O and t_N are compared. From Table 5.7, $\text{MSE}(t_O) = 34786.61$ and $\text{MSE}(t_N) = 34721.49$ with corresponding PRE's of 100.0% and 100.16%. These PRE's indicate that if stratification were to be used instead of SRSWOR, there would be an increase in precision of the estimator by 0.16%, an observation consistent with the finding of Ngesa et al. (2012).

The results in Table 5.7 further indicate that both \hat{Y}_{MD} and \hat{Y}_{MDC} perform better than t_O since the respective PRE's indicates an improved precision by 3.54% and 0.1%. Still this is an indication that stratification improves efficiency of estimators. However, when performances of both \hat{Y}_{MD} and \hat{Y}_{MDC} are compared with that of t_N , a different observation is made. Based on the PRE's, it can be seen that t_N is slightly more efficient than \hat{Y}_{MDC} but less efficient than \hat{Y}_{MD} . On the question of whether to use separate multivariate ratio estimator or combined multivariate ratio estimator, respective MSE's of \hat{Y}_{MD} and \hat{Y}_{MDC} are compared as shown below

$$\text{PRE}(\hat{Y}_{MD}) = \left(\frac{\text{MSE}(\hat{Y}_{MDC})}{\text{MSE}(\hat{Y}_{MD})} \times 100 \right) - 100 = 3.5\% \quad (5.80)$$

From equation (5.80), it is clear that when separate multivariate ratio estimator is used instead of a combined multivariate ratio estimator, there is a gain in precision by 3.5%. Using the PRE's and the population total estimates of the multivariate estimators, it can be conclude that Y_{MD} is most efficient and has the least absolute deviation from the actual population total.

Though based on the results in Table 5.7, it is observed that \hat{Y}_{MD} performs better than the other multivariate ratio estimators, the estimates, however, depend on the choice of the weights. From equation (5.70), the expressions of the optimal values for weights W_{cj1} and W_{cj2} for $p = 2$ is given by

$$W_{cj1} = \frac{V_{cj2}}{(V_{cj1} + V_{cj2})} \text{ and } W_{cj2} = \frac{V_{cj1}}{(V_{cj1} + V_{cj2})}$$

For Olkin's estimator (1958), response groupings are used as two different strata such that for $j = 1$,

$$W_{11} = \frac{48414.51}{48414.51 + 60543.53} = 0.44$$

and

$$W_{12} = \frac{60543.53}{48414.51 + 60543.53} = 0.56$$

or equivalently,

$$W_{12} = 1 - 0.44 = 0.56.$$

Table 5.8: Univariate Variances and Optimal Weights

Estimator	Stratum	Resp. group	Univariate Variances		Optimal Weights	
			l=1	l=2	W_{cj1}	W_{cj2}
t_O	-	j=1	48414.51	60543.53	0.44	0.56
		j=2	13642.42	11741.06	0.54	0.46
t_N	c=1	j=1	23291.30	36836.1	0.39	0.61
		j=2	6362.19	7363.05	0.46	0.54
	c=2	j=1	15683.60	15353.80	0.51	0.49
		j=2	2504.44	2833.22	0.47	0.53
	c=3	j=1	8392.70	8335.80	0.50	0.50
		j=2	1885.05	1948.68	0.49	0.51
\hat{Y}_{MD}	c=1	j=1	21276.40	37777.13	0.36	0.64
		j=2	6527.39	8831.62	0.42	0.58
	c=2	j=1	15208.02	14479.42	0.51	0.49
		j=2	2291.84	2803.02	0.45	0.55
	c=3	j=1	8662.73	8224.63	0.51	0.49
		j=2	2066.14	1971.30	0.51	0.49
\hat{Y}_{MDC}	c=1	j=1	22647.60	34013.20	0.40	0.60
		j=2	6630.97	7987.40	0.45	0.55
	c=2	j=1	15002.10	15032.30	0.50	0.50
		j=2	2100.27	2662.76	0.44	0.56
	c=3	j=1	9083.28	9073.94	0.50	0.50
		j=2	2617.08	2493.46	0.51	0.49

Using this formula, other weights for other multivariate ratio estimators are computed as shown in Table 5.8.

From Table 5.8, the column for $l = 1$ represents univariate variances using the first component (X_1) of the auxiliary random vector, while the second column for $l = 2$ represents univariate variances using the second component. In Table 5.2, values for the weights for each stratum in each response group are randomly generated, while for SRSWOR (Olkin's multivariate estimation), the averages of the randomly generated weights are used. On the other hand, under minimum variance, the optimal values for the weights in each response group and stratum are used. For this reason, different values for the weights are used for each multivariate ratio estimator in each stratum and response group.

Using these optimal weights shown in Table 5.8, a summary of the optimal MSE's is obtained as shown in Table 5.9. Only MSE's and PRE's of the multivariate estimators are considered.

Table 5.9: Minimum Variance under Optimal Weights

	Stratum 1		Stratum 2		Stratum 3		Total
	j=1	j=2	j=1	j=2	j=1	j=2	
MSE(t_O)	$j = 1, \text{MSE}(t_{O1}) = 28252.15$		$j = 2, \text{MSE}(t_{O2}) = 6452, 69$				34704.83
MSE(t_N)	17320.26	3486.05	7761.98	1349.60	4182.27	959.23	35059.39
MSE(\hat{Y}_{MD})	18221.36	4099.02	7435.27	1312.18	4230.36	1011.03	36309.22
MSE(\hat{Y}_{MDC})	15875.07	3748.99	7508.62	1240.58	4539.31	1279.88	34198.45
PRE(t_O)							100.00%
PRE(t_N)							98.99%
PRE(\hat{Y}_{MD})							95.58%
PRE(\hat{Y}_{MDC})							101.50%

Table 5.9 shows that \hat{Y}_{MDC} has the least variance while \hat{Y}_{MD} has the highest MSE. This implies that the combined multivariate unbiased ratio estimator is most efficient among the multivariate ratio estimators. The PRE'S For t_N and \hat{Y}_{MD} is an indication that the Olkin's estimator performs better than these two. For \hat{Y}_{MDC} , the gain in precision over Olkin's multivariate estimator (1958) is 1.5%. Based on this observation, we can thus conclude that \hat{Y}_{MD} performs better when we use randomly generated weights, while Y_{MDC} performs better when we use optimal values of the weights.

5.6 Chapter Summary

In this chapter, regression-based unbiased ratio estimators for finite population total in stratified random sampling under non-response is constructed. The problem of non-response in the study variable has been addressed using the Hansen-Hurwitz sub-sampling method. Using an unbiased ratio estimator, both univariate and multivariate regression-based estimators have been successfully constructed. Asymptotic properties of the suggested regression-based unbiased ratio estimators, such as biasness, variance and optimality conditions, have been studied. Using simulated data, performance of the

proposed estimators have shown improved efficiency relative to the known regression-based estimators in literature. Simulation results have supported the theoretical proofs in both chapter four and chapter five. Conclusively, the suggested ratio-type estimators perform better than the known ratio-estimators in literature.

6. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

6.1 *Introduction*

This chapter summarizes the work done, conclusions and recommendations for further research.

6.2 *Summary*

In sample surveys, use of ratio estimators in estimation procedures has been extensively adopted by several researchers. This has been done using one or more auxiliary variables or attributes. As a result, several ratio-type estimators have been constructed. These estimators have, however, suffered a common drawback that the estimators are biased, though the bias becomes negligible as the sample size increases. Little effort has been done on how to minimize bias of ratio estimators while retaining their essential estimators. However, using the bias of the usual ratio estimator, an unbiased ratio estimator has been constructed. In this study, an unbiased ratio estimator for finite population total in stratified random sampling under non-response has been constructed.

In this study, the main objective of construction of an unbiased ratio estimator for finite population total in stratified random sampling under non-response has been motivated by the fact that standard ratio estimators are biased and thus, the need to develop unbiased ratio estimators. Therefore, unbiased ratio estimator in stratified random sampling using both separate and combined ratios has been developed, while considering both univariate and multivariate estimation procedures. The regression form of the constructed unbiased ratio estimator has also been obtained. In each case, asymptotic properties of the constructed estimators have been studied. Further, optimum conditions under which the constructed estimators perform best have

been established. Using simulated data, asymptotic properties of these unbiased ratio-type estimators have been verified.

In simulation, a hypothetical population of 300 units with three strata of stratum sizes 167, 94 and 39 has been considered. Random values have been generated from normal population by changing the parameters for each variable in each stratum and a linear model between the study variable and the auxiliary variable was then fit. Simple random samples were then obtained from each stratum without replacement and various sample statistics computed. For regression ratio estimation, a predetermined regression model of the study variable on the auxiliary variable has been assumed, while for multivariate unbiased ratio estimation, a two-component auxiliary random vector has been considered. Performance of constructed unbiased ratio-type estimators have been evaluated using their respective percent relative efficiency.

From the theoretical and simulation results, it has been observed that, due to the observed least MSE's, the suggested estimators perform better than the usual ratio estimator under SRSWOR. This improved efficiency has been observed for both separate and combined unbiased ratio estimators. This finding is consistent with the knowledge that stratification improves the level of precision of estimators. Moreover, when compared using stratified random sampling, the suggested unbiased ratio estimator has still been observed to perform better than the usual ratio estimator. Among a class of unbiased ratio estimators, the results indicate that the unbiased ratio estimator is a best linear unbiased estimator since it is not only unbiased but has the least variance among a class of unbiased linear estimators for finite population total.

While comparing separate and combined unbiased ratio estimators, it has been noted that superiority of either separate or combined ratio estimators over the other depends on the absolute difference between the stratum mean ratios and the overall mean ratio in each response group. However, despite this observation, the results further indicate that the unbiased regression ratio estimator performs better than both separate and combined unbiased ratio estimators. This is based on the corresponding PRE's. In particular, the regression form of the unbiased ratio estimator has a minimum variance

only when the covariance is the product of the regression coefficient and variance of the auxiliary variable, an observation consistent with the known knowledge on regression estimators.

In multivariate ratio estimation, it has been observed that the estimates of population parameters depend on the choice of the weights for each component of the auxiliary random vector. In multivariate ratio estimation, weights can be randomly generated such that the sum equals to 1 (one) or can be computed using the expressions for the optimal weights. Using randomly generated values for weights in each stratum, it was observed that the separate unbiased ratio estimator performs better than the combined unbiased ratio estimator. This observation is, however, reversed when optimal values of the weights are used. That is, under optimal weights, the combined unbiased ratio estimator performs better than the separate unbiased ratio estimator. All these observation have been based on large sample approximations and proportional allocation of sample sizes in each stratum.

6.3 Conclusion

In this study, the general objective was to construct an unbiased ratio estimator for finite population total in stratified random sampling under non-response. Specifically, the study aimed at constructing an unbiased ratio estimator in stratified random sampling under non-response. Also, the study aimed at deriving regression and multivariate forms of the constructed estimator. Moreover, the study aimed at carrying out a simulated study to compare performance of the constructed unbiased ratio estimators. To achieve these specific objectives, the assumption of large samples, proportional allocation of sample sizes in each stratum and non-response only on the response variable has been made. From the results, it can be concluded that using Hansen-Hurwitz sub-sampling technique, the constructed unbiased ratio estimator performs better than the usual ratio estimator constructed under SRSWOR and under stratified random sampling. Nevertheless, the choice between separate unbiased ratio estimator over combined unbiased ratio estimator, or vice versa, depends on the absolute difference between the mean ratios in each stratum and the overall mean ratio.

Efficiency of a ratio estimator is improved when there is a perfect linear relationship between the response variable and the auxiliary variable. This has been confirmed by the properties of the regression unbiased ratio estimator. From the properties of the regression ratio-type estimator, it can be concluded that regression estimation yields estimators with high levels of precision, especially when the regression line of the response variable on the auxiliary variable passes through the origin. Further, under optimal conditions, precision of the constructed regression unbiased ratio estimator is improved when the correlation coefficient in each stratum and response group asymptotically tends to one.

Though the general form of the multivariate ratio estimator involves expressing finite population total as a linear function of the totals obtained using each component of the auxiliary random vector pre-multiplied by some weights, it has been observed that the choice of the values of the weights is also vital in multivariate ratio estimation procedures. That is, from the results, it can be concluded that multivariate ratio estimates of various population parameters and the corresponding MSE's of different multivariate ratio estimators is highly dictated by the values of the weights. Therefore, to evaluate performance of any multivariate ratio estimator, only optimal values of the weights should be used.

6.4 Contributions of the Study to Knowledge

This study focused on how to estimate finite population total in stratified random sampling under non-response using an improved ratio estimator. This study has shown how sub-sampling technique can be used to address the problem of non-response, especially in stratified random sampling technique. In particular, this study has shown how an unbiased ratio estimator can be constructed using separate, combined, regression and multivariate ratio estimation procedures. This is a major contribution to literature development since in each case, the asymptotic properties have not only been studied, but have also been verified using simulated data. Moreover, in each case, optimality conditions have been studied, an area that has not been extensively studied. The findings of this study are, therefore, crucial in literature development on ratio estimation problems involving more than

one auxiliary variable and where there is non-response.

6.5 *Recommendations for Further Research*

From the results of this study, we recommend that the following areas still require further research:

- (i) Construction of an unbiased ratio estimator under non-response in both response variable and auxiliary variable(s).
- (ii) Construction of the regression form of the unbiased ratio estimator when there is no perfect relationship between the survey variable and auxiliary variable(s)
- (iii) Construction of an unbiased ratio estimator in the case of auxiliary attribute(s) under non-response.
- (iv) Construction of a multivariate form of unbiased ratio estimator when both auxiliary variable and response variable are random vectors.
- (v) Construction of a multivariate form of unbiased ratio estimator when $p \geq 3$.
- (vi) Construction of a multivariate form of unbiased ratio estimator when $\text{Cov}(Y_{cjl}, Y_{cjl'}) \neq 0$.

BIBLIOGRAPHY

- [1] Asghar, A., Sanaullah, A. & Hanif, M. (2017). A Multivariate Regression-cum- Exponential Estimator for Population Variance Vector in Two Phase Sampling. *Journal of King Saud University – Science*. dx.doi.org/10.1016/j.jksus.2017.01.010
- [2] Chakrabarty, R. P. (1979). Some Ratio Estimators. *Journal of the Indian Society of Agricultural Statistics*, 31(1), 49–57.
- [3] Chaudhary, M. K. & Kumar, A. (2015). Estimating the Population Mean in Stratified Random Sampling Using Two-Phase Sampling in the Presence of Non-Response. *World Applied Sciences Journal*, 33 (6), 874-882.
- [4] Chaudhary M. K., Malik S., Singh J. & Singh R. (2013). A General Family of Estimators for Estimating Population Mean in Systematic Sampling Using Auxiliary Information in the Presence of Missing Observations. *Jour. Raj. Stat. Assoc.*, 1(2), 1-8.
- [5] Cochran W. G. (1977). *Sampling Techniques*. 3rd Edition. New York, John Wiley.
- [6] Daroga, S. & Chaudhary F. (2002). *Theory and Analysis of Sample Survey Designs*. New Delhi, New Age International (P) Limited Publishers.
- [7] Dorofeev, S. & grant, P. (2006). *Statistics for Real Life Sample Surveys. Non-Simple Random Samples and Weighted Data*. Cambridge University Press, New York.
- [8] Durbin, J. (1959). A Note on the Application of Quenouille’s Method of Bias Reduction to the Estimation of Ratios. *Biometrika*, 46 (3 and 4), 477-480.

- [9] Goodman, L. A. & Hartley, H. O. (1958). The Precision of Unbiased Ratio-Type Estimators. *Journal of the American Statistical Association*, 53, 491-508.
- [10] Hamad, N., Haider, N. & Hanif, M. (2013). A Regression Type Estimator with Two Auxiliary Variables for Two-Phase Sampling. *Open Journal of Statistics.*, 3, 74-78.
- [11] Hanif, M., Hamad, N. & Shahbaz, M. Q. (2009). A Modified Regression Type Estimator in Survey Sampling. *World Applied Sciences Journal.*, 7(12), 1559- 1561.
- [12] Hanif, M., Shahbaz, M. Q. & Ahmad, Z. (2010). Some Improved Estimators in Multiphase Sampling. *Pakistan Journal of Statistics.*, 26(1), 195 – 202.
- [13] Hansen, M. H. & Hurwitz, W. N. (1943). On the Theory of Sampling from Finite Populations. *The Annals of Mathematical Statistics.*, 14(4), 333-362.
- [14] Hansen M. H. & Hurwitz W. N. (1946). The Problem of Non Response in Sample Surveys. *Journal of the American Statistical Association*, 41, 517-529.
- [15] Hartley, H. O., & Ross, A. (1954). Unbiased Ratio Estimators. *Nature*, 174, 270-1.
- [16] Ismail, M., Shahbaz, M. Q. & Hanif, M. (2011). A General Class of Estimator of Population Mean in Presence of Non-Response. *Pak. J. Statist.*, 27(4), 467-476.
- [17] Kadilar, C. & Cingi, H. (2004). Ratio Estimators in Simple Random Sampling. *Applied Mathematics and Computation.*, 151, 893-902.
- [18] Kadilar, C. & Cingi, H. (2006). An Improvement in Estimating the Population Mean by Using the Correlation Coefficient. *Hacettepe Journal of Mathematics and Statistics.*, 35, 103-109.
- [19] Khare, B. B. & Rehman, H. U. (2015). Improved Ratio in Regression Type Estimator for Population Mean Using Known Coefficient of Variation

- of the Study Character in The Presence of Non-Response. *International Journal of Technology Innovations and Research (IJTIR)*., 14, 1-7.
- [20] Khoshnevisan, M., Singh, R., Chauhan, P., Sawan, N. & Smarandache, F. (2007). A General Family of Estimators for Estimating Population Mean Using Known Value of Some Population Parameters. *Far East journal of Theoretical Statistics*, 22, 181-191.
- [21] Koop, J. C. (1951). A Note on the Bias of the Ratio Estimate. *Statistical Institute Bulletin*, 33(2), 141-6.
- [22] Kumar, S. (2012). Utilization of Some Known Population Parameters for Estimating Population Mean in Presence of Non-Response, *Pak.j.stat.oper.res.*, 8(2), 233-244.
- [23] Lone, H. A. & Tailor, R. (2015). A Family of Estimators for Estimating Population Variance Using Auxiliary Information in Sample Survey. *Pak.j.stat.oper.res.* 11 (2), 213-220.
- [24] Mangat, S. N. & Singh, R. (1996). *Elements of Surveys Sampling*. Springer-Science and Business Media, India.
- [25] Murthy, M. N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, India.
- [26] Mohanty, S. (1967). Combination of Regression and Ratio Estimate. *Journal of Indian Statistical Association.*, 5. 16-19.
- [27] Montanari, G. E. (1998). On Regression Estimation of Finite Population Means. *Survey Methodology.*, 24(1), 69-77.
- [28] Mukerjee, R., Rao, T. J. & Vijayan, K. (1987). Regression Type Estimators Using Multiple Auxiliary Information. *Australian Journal of Statistics.*, 29(3), 244-254.
- [29] Ngesa, O. O., Orwa, G. O., Otieno, R. O. & Murray, H. M. (2012). Multivariate Ratio Estimator of the Population Total under Stratified Random Sampling. *Open Journal of Statistics*, 2, 300-304.

- [30] Olayiwola O. M., Popoola R. A. & Bisira H. O. (2016). A Modified Regression Estimator For Double Sampling. *Annals. Computer Science Series.*, 14th Tome 1st Fasc., 44-49.
- [31] Olkin, I. (1956). Multivariate Ratio Estimation for Finite Populations. *Biometrika*, 45(1-2), 154-165.
- [32] Oyoo D. & Ouma C. (2014). Estimation of Population Total in the Presence of Missing Values using a Modified Murthy's Estimator and the weight Adjustment Technique. *American Journal of Applied Mathematics and Statistics*. 2(3), 163-167.
- [33] Oyoo, D., Manene, M., Ouma, C. & Muhua, G. (2019). Modified Ratio Estimator of Finite Population Total in Stratified Random Sampling under Non-Response. *Mathematical Theory and Modeling*, 9(7), 74-88.
- [34] Pfefferman, D. & Rao, C. R. (Eds.) (2009). *Handbook of Statistics 29A: Design, Methods and Applications*. Elsevier Publications, North Holland.
- [35] Rao, J. N. K. (1965). A Note on Estimation of Ratios by Quenouille's Method. *Biometrika*, 52, 647-9.
- [36] Ray, S. K. & Sahai, A. (1980). Efficient Families of Ratio and Product Type Estimators. *Biometrika*, 67(1), 211-215.
- [37] Robson, D. S. (1957). Applications of Multivariate Polykays to the Theory of Unbiased Ratio-Type Estimation. *Journal of the American Statistical Association*, 52, 511-522.
- [38] Roy, D. C. (2003). A Regression Type Estimator in Two Phase Sampling Using Two Auxiliary Variables. *Pakistan Journal of Statistics.*, 19(3), 281-290.
- [39] Sahoo, J., Sahoo L. N. & Mohanty, S. (1993). A Regression Approach to Estimation in Two Phase Sampling Using Two Auxiliary Variables. *Current Sciences.*, 65, 73-75.
- [40] Sampath, S. (2001). *Sampling Theory and Methods*. Narosa Publishing House, New Delhi, India.

- [41] Singh, R. & Smarandache, F. (Eds.) (2013). On Improvement In Estimating Population Parameter(s) Using Auxiliary Information. Educational Publishing (Columbus) & Journal of Matter Regularity (Beijing), USA - China, 25 - 41.
- [42] Solanki, R.S., Singh, H. P. & Rathour, A. (2012). An Alternative Estimator for Estimating the Finite Population Mean Using Auxiliary Information in Sample Surveys. ISRN Probability and Statistics doi:10.5402/2012/657682.
- [43] Tin, M. (1965). Comparisons of some ratio estimators. *Journal of American Statistical Association.*, 60(309), 294-307.
- [44] Raj, D. (1965). On a Method of Using Multi-Auxiliary Information in Sample Surveys. *Journal of the American Statistical Association.*, 60(309), 270-277.
- [45] Rao, P. S. R. S. (1986). Ratio Estimation with sub sampling the non-respondents. *Survey Methodology*, 12(2), 217-230.
- [46] Reddy, V. N. (1973). On ratio and product method of estimation. *Sankhya*, B, 35, 307-317.
- [47] Robson, D. S. (1957). Application of Multivariate Polykeys to the Theory of Unbiased Ratio Type Estimators. *Journal of the American Statistical Association.*, 52(280), 511-522.
- [48] Samiuddin, M. & Hanif, M. (2007). Estimation of Population Mean in Single Phase and Two-Phase Sampling with or without Additional Information. *Pakistan Journal of Statistics.*, 23(2), 99-118.
- [49] Shabbir, J. & Saghir, A. (2012). Estimation of Finite Population Mean in Stratified Random Sampling Using Auxiliary Attribute(s) under Non-Response. *Pak.j.stat.oper.res*, 8(1), 73-82.
- [50] Singh, R. and Malik, S. (2014). Improved estimation of population variance using information on auxiliary attribute in simple random sampling. *Applied Mathematics and Computation*, 235, 43-49.
- [51] Srivastava, S. K. (1967). An estimator using auxiliary information in sample surveys. *Calcutta Statistical Association Bulletin*, 16, 121-132.

- [52] Srivastava, S. K. (1971). A Generalized Estimator for the Mean of a Finite Population Using Multi Auxiliary Information. *Journal of the American Statistical Association.*, 66(334), 404-407.
- [53] Subramani, J., Kumarapandiyan, G. & Balamurali, S. (2014). Some Modified Linear Regression Type Ratio Estimators for Estimation of Population Mean Using Known Parameters of an Auxiliary Variable. *Journal of Buildings Research.*, 1(1), 28-42.
- [54] Sukhatme, B. V. (1962). Some Ratio Type Estimators in Two- Phase Sampling. *Journal of the American Statistical Association.*, 57(299), 628-632.
- [55] Yan, Z. & Tian, B. (2010). Ratio Method to The Mean Estimation Using Coefficient of Skewness of Auxiliary Variable. *ICICA 2010, Part II.*, 106, 103–110.
- [56] Walsh, J. E. (1970). Generalization of ratio estimates for population total. *Sankhya, A*, 32, 99-106.
- [57] Zaman, T. & Yilmaz, S. (2017). Review of Some Ratio Estimators in Stratified Random Sampling. *J. Math. Comput. Sci.*, 7 (2), 364-374.

Appendix A

SIMULATION CODE

```
> Consider a hypothetical population of size N=300 with 3 strata
> p<-runif(3)
> q<-p/sum(p)
> N'<-300*q
> N<-round(N', digits=0)
> Stratum Population Sizes<- 94 39 167
> # Use Krejcie-Morgan-Sample-Size-Table to get overall sample size as
170 and proportional allocation to get stratum sample sizes

> A11<-rnorm(167,35,12)
> B11<-rnorm(167,45,14.5)
> Y1<-round(A11,digits=2) # response variable in stratum 1
> X1<-round(B11,digits=2) # auxiliary variable in stratum 1
> Stratum1<-as.data.frame(cbind(X1,Y1))
> lm1<-lm(Y1~X1)

> A22<-rnorm(94,35,11)
> B22<-rnorm(94,40,12)
> Y2<-round(A22,digits=2) # response variable in stratum 2
> X2<-round(B22,digits=2) # auxiliary variable in stratum 2
> Stratum2<-as.data.frame(cbind(X2,Y2))
> lm2<-lm(Y2~X2)

> A33<-rnorm(39,48,14)
> B33<-rnorm(39,32,16)
> Y3<-round(A33,digits=2) # response variable in stratum 3
> X3<-round(B33,digits=2) # auxiliary variable in stratum 3
> Stratum3<-as.data.frame(cbind(X3,Y3))
```

```
> lm3<-lm(Y3~X3)
> # Assume non-response rate of 20% in each stratum
> # Partition each stratum into the two response groups in each stratum
>
>
>
> S11id<-sample(1:133,75,replace=FALSE,prob=NULL)
> S12id<-sample(1:34,19,replace=FALSE,prob=NULL)
> m1id<-sample(S12id,16,replace=FALSE,prob=NULL)
> S21id<-sample(1:75,42,replace=FALSE,prob=NULL)
> S22id<-sample(1:19,11,replace=FALSE,prob=NULL)
> m2id<-sample(S22id,10,replace=FALSE,prob=NULL)
> S31id<-sample(1:31,18,replace=FALSE,prob=NULL)
> S32id<-sample(1:8,5,replace=FALSE,prob=NULL)
> m3id<-sample(S32id,4,replace=FALSE,prob=NULL)
```