ZHANG, QI, Ph.D. Mean Estimation of Sensitive Variables under Measurement Errors and Non-Response. (2020)
Directed by Dr. Sat Gupta. 118 pp.

This study mainly consists of three important issues we face in survey sampling: social desirability bias, measurement errors, and non-response. In this dissertation, we study the mean estimation of a sensitive variable under measurement errors and non-response. We propose a generalized mean estimator, then discuss the bias and the mean square error (MSE) of this estimator and present the comparisons with other estimators under the measurement errors and non-response using optional RRT model (ORRT). We also study the performance of the proposed estimator under the same situations using stratified random sampling. Simulation studies are also conducted to verify the theoretical results. Both the theoretical and empirical results show that the generalized mean estimator is more efficient than the ordinary RRT estimator that does not utilize the auxiliary variable, and the ratio estimator which is one of the commonly used mean estimator.

MEAN ESTIMATION OF SENSITIVE VARIABLES UNDER MEASUREMENT ERRORS AND NON-RESPONSE

by

Qi Zhang

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2020

Approved by

____
Committee Chair

APPROVAL PAGE

This dissertation written by Qi Zhang has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair ______________________________
Sat Gupta

Committee Members ______________________________
Scott Richter

____
Jianping Sun

____
Somya Mohanty

____
Rakhi Singh

____
Date of Acceptance by Committee

____
Date of Final Oral Examination

ACKNOWLEDGMENTS

This thesis would not have been possible without the help of so many people in so many ways.

Most of all, I would like to thank my mentor, Dr. Sat Gupta, for motivating me to work on this topic, helping me to deeply understand and giving substantial advice regarding the topic of my research. I would also like to express my appreciations for his understanding, patience, encouragement and for pushing me further than I thought I could go. Dr. Gupta is an excellent mentor, he always inspired me with a desire to learn. He was in influential in guiding me, while boosting my confidence as well. It was a pleasure working with him.

Next I would like to thank Dr. Scott Richter, Dr. Jianping Sun, Dr. Somya Mohanty, and Dr. Rakhi Singh for the assistance and direction they have provided me. I would also like to express my most sincere appreciation to Sadia Khalil of Lahore College for Women University also who is my co-author on several papers and who was always available for discussions and guidance.

Also, I would like to express my gratitude to all my professors throughout my college career at UNCG who have assisted me in becoming a better statistician. Also, I am appreciative of the Department of Mathematics and Statistics which provided me the financial support and an opportunity to obtain my PhD degree.

Last but not least, I would like to express my sincere appreciation to all my friends and family for helping me survive all the stress, and not letting me give up.

TABLE OF CONTENTS

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION AND BACKGROUND

In statistical studies, one of the fundamental goals is to estimate the true value of population parameters. Unfortunately, collecting data from every member of the population would be too expensive or time-consuming if the population is large. Instead of conducting a census, we can collect data from a sample and use the sample statistics to make inferences about the target population. However, sometimes a sample cannot represent the population accurately due to several sampling or non-sampling errors. Sampling error is an error caused by working with a part of the population and not the whole population. Most of the time it can be reduced by increasing the sample size. Non-sampling errors can be attributed to several problems including respondent mistakes, measurement errors and non-response, etc. Therefore, the inferences will be reasonable only if the sample truly represents the population and the responses collected from the sample are accurate. Otherwise the sample is biased and conclusion from the study are not trustworthy.

There are many sampling methods, such as the simple random sampling (SRS), cluster sampling, and stratified random sampling, etc. In order to have a representative sample, we could use different sampling methods depending on the situation. SRS is the basic sampling technique where each member of the population has an equal chance of being selected in the sample. But sometimes sub-populations within an overall population vary significantly, and it would be better to sample each sub-population independently. This refers to stratified random sampling.

We also have different types of survey methods that are often used, such as email surveys, phone surveys and personal interview surveys. Email and phone surveys are relatively cheaper but usually have high non-response rate. The non-response issue may cause some participation bias. For example, people who feel strongly about an issue may be more likely to participate, and their opinion may not represent the opinion of the whole population. People are unlikely to reject a personal interview survey compared to the other two methods but it costs more. Additionally, if the survey question is sensitive, a personal face-to-face interview may cause social desirability response bias. For example, if a survey question asks "What is your salary?" or "Have you ever used illegal drugs?", most people would want to present themselves in a socially desirable light, therefore their responses may be biased towards what they feel is socially desirable. In addition to participation bias and social desirability response bias, there are some other non-sampling errors, such as measurement errors occurring due to definition differences or misunderstandings which will also affect parameter estimation. Therefore, dealing with these problems is very important when we estimate the population parameters of a sensitive variable.

This dissertation will consist of three important issues we face in survey sampling: social desirability bias, measurement errors, and non-response. We will study mean estimation for a sensitive variable in the presence of such issues using both SRS and stratified random sampling. In Section I.1, we will introduce some RRT models which improve efficiency of the mean estimation if a survey involves sensitive questions. In Section I.2, we will demonstrate how the Hansen and Hurwitz two-phase sampling technique works if non-response exists. In Sections I.3 and I.4, the basic ideas of measurement errors and two common sampling methods will be briefly described. An outline of the dissertation will be presented in Section I.5.

I.1. Randomized Response Technique Models

Randomized response technique (RRT) is an important method to prevent or reduce social desirability response bias and is widely used in survey interviews. The first RRT model was proposed by Warner in 1965 [84]. It was modified later by many researchers including Greenberg et al. (1969) [15], Eichhorn (1983) [10], Gupta et al. (2002) [17], Gupta et al. (2010) [19], and Sousa et al. (2010) [77] etc. It allows respondents to answer sensitive questions more comfortably and provides more accurate estimates. RRT models have been used in many field surveys, such as Kerkvliet et al. (1994) [32], Gill et al. (2013) [14], Chhabra et al. (2016) [06], Chen et al. (2014) [05], and Geng et al. (2016) [13]. Several RRT models will be described in detail in this section, but the optional RRT model (ORRT) will be the main focus in this dissertation.

I.1.1 Warner's Binary RRT Model (1965)

In 1965, Warner [84] proposed the first binary RRT model to estimate the prevalence of a sensitive characteristic in a population. It increases response rate, and also makes the respondent feel more comfortable in answering survey questions truthfully and reduces social desirability response bias. Warner's Binary RRT model will be illustrated below by an example.

Suppose we are interested in estimating what proportion of college students have a sexually transmitted disease (STD). A randomization device, for instance a deck of cards that contains two questions (or statements), will be used in this survey to divide the sample into two groups. The two statements may be:

1. I have been told by a healthcare professional that I have STD.

2. I have never been told by a healthcare professional that I have STD.

A known proportion p of the cards in the deck contain Statement 1, and the remaining cards contain Statement 2. A simple random sample of n respondents is drawn from the population, and each subject is asked to pick a card from the deck and provide a "yes" or a "no" response to the statement on the card. Among these n subjects, let there be n_1 respondents who answer "yes". A "yes" response does not mean this person has STD; there is another possibility that the person may have picked the second statement. The same is true for a "no" response. In this case, the interviewer does not know which statement the respondent picked. And the respondent is more likely to provide a true response since his/her privacy is assured.

Let π be the true probability of a subject having an STD in the population, and p_y be the probability of a "yes" responses. Then,

$$p_y = p\pi + (1-p)(1-\pi). \tag{I.1}$$

The estimate of π is then given by

$$\hat{\pi} = \frac{\hat{p}_y - (1-p)}{2p-1} = \frac{\frac{n_1}{n} - (1-p)}{2p-1}. \tag{I.2}$$

The variance of this estimator under simple random sampling with replacement is given by

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}. \tag{I.3}$$

In order to minimize the variance, a large sample size n should be chosen and the proportion (p) of Statement 1 should be closer to 0 or 1.

I.1.2 Warner's (1971) and Pollock & Beck's (1976) Quantitative RRT Model

The estimator proposed in 1965 can be used to estimate binary variables, but many times the question of interest is a quantitative one. Warner [85] modified the RRT model for quantitative cases in 1971. This was further expanded by Pollock and Bek (1971) [53]. We use another example to illustrate this model.

Suppose we are interested in estimating how many sexual partners a college student had in the last 3 months. Instead of creating cards with two questions in the deck, we make cards with random numbers from a pre-assigned distribution, preferably with mean zero . The respondents are asked to pick a card, and add the number on the card to their true answer. Then they report a number, which is the sum of the true answer and the random number they picked. Let Y be the sensitive variable with unknown mean μ_y and unknown variance σ_y^2, and S be the scrambling variable (independent of Y) with known mean μ_s and known variance σ_s^2. Also let Z be the reported response. Then

$$Z = Y + S. \tag{I.4}$$

The expected response is given by

$$E(Z) = E(Y) + E(S). \tag{I.5}$$

This leads to an unbiased estimator of the mean of the sensitive variable Y is given by

$$\hat{\mu}_y = \bar{z} - \mu_s. \tag{I.6}$$

or simply $\bar{z}$ if μ_s=0.

The variance of $\hat{\mu}_y$ is given by

$$Var(\hat{\mu}_y) = Var(\bar{Z}) = \frac{Var(Z)}{n} = \frac{\sigma_y^2 + \sigma_s^2}{n} = \frac{\sigma_y^2}{n} + \frac{\sigma_s^2}{n}, \tag{I.7}$$

where $\frac{\mu_s^2}{n}$ is the penalty for using the RRT model.

I.1.3 Eichhorn and Hayre's Multiplicative RRT Model (1983)

Eichhorn and Hayre [10] introduced a multiplicative RRT model. Instead of adding a random number to the true response, the respondent needs to multiply the true response by a randomly selected number from a known distribution and divide by the mean of the scrambling variable.

The reported response is given by $Z = YS/\mu_s$, where $\mu_s = E(S)$. Usually μ_s is chosen to be 1. This leads to the unbiased estimator

$$\hat{\mu}_y = \bar{z}, \tag{I.8}$$

where $\bar{z}$ is the sample mean of the reported responses. The variance of $\hat{\mu}_y$ is given by

$$Var(\hat{\mu}_y) = \frac{1}{n}[\sigma_y^2 + \frac{\sigma_s^2(\sigma_y^2 + \mu_y^2)}{\mu_s^2}]. \tag{I.9}$$

I.1.4 Gutpa et al. Optional RRT Model (2002)

In the above models, every respondent is forced to provide a scrambled response. However, researchers have realized that a question may be sensitive for one respondent, but not sensitive for another. In order to make the survey results more accurate, Gupta et al. (2002) [17] modified the Eichhorn and Hayre (1983) [10] multiplicative scrambling RRT model, and introduced an Optional RRT (ORRT) model that allows researchers to estimate not only the mean of the variable of interest, but also the sensitivity level W (the proportion of subjects in the population who consider the question sensitive).

In this model, if respondents feel the question is sensitive, they will provide a scrambled response YS. Otherwise, the respondents will answer the sensitive survey question directly and provide the true response Y. In this model, we assume that both Y and S are positive valued random variables and that the mean of the scrambling variable $\mu_s = 1$ and the variance σ_s^2. Under this model, the reported response Z is given by

$$Z = \begin{cases} Y & \text{with probability 1-W} \\ YS & \text{with probability W.} \end{cases} \tag{I.10}$$

The expected value of Z is given by

$$E(Z) = E(Y)(1-W) + E(YS)W = \mu_y(1-W) + \mu_y\mu_s W = \mu_y, \tag{I.11}$$

and the variance of this unbiased estimator of the population mean μ is given by

$$Var(\hat{\mu}_y) = \frac{1}{n}[\sigma_y^2 + W\sigma_s^2(\sigma_y^2 + \mu_y^2)]. \tag{I.12}$$

Note that $Var(\hat{\mu}_y)$ increases with W, and hence there is gain in efficiency compared to the non-optional model where W=1. Gupta et al (2002) [17] gave an estimator for the sensitivity level W which is given by

$$\hat{W} = \frac{\frac{1}{n}\sum_{i=1}^{n} log(Z_i) - log(\frac{1}{n}\sum_{i=1}^{n} Z_i)}{\delta}, \tag{I.13}$$

where δ=E[log(S)].

I.1.5 Gupta et al. Optional Additive RRT Model (2010)

The multiplicative scrambling compromises respondent anonymity. For example, if the respondent's true response is zero, no matter what scrambling number s/he chooses, the reported response will be zero. In this case, a non-zero response means the respondent has some degree of the sensitive characteristic. Another shortcoming of the multiplicative scrambling model is that some respondents may not like to multiply or may not know how to multiply the scrambling variable. The respondents still provide untruthful response. Singh et al. (1996) [64] showed that this case is more dangerous than not using the scrambled response. In order to deal with these problems and also to estimate the sensitively level W without using any approximations, Gupta et al. (2010) [19] proposed an additive ORRT model using a split-sample approach.

The split-sample approach means that we split the sample into two subgroups. One group of respondents uses a scrambling variable S_1, and the other group uses a different scrambling variable S_2. Since the multiplicative scrambling method compromises respondent anonymity, the scrambling method used is additive scrambling.

Again, let Y be a sensitive variable with mean μ_y, S_i $(i = 1, 2)$ be scrambling variable (independent of Y) with mean μ_{s_i} $(i = 1, 2)$ and variance $\sigma^2_{s_i}$ $(i = 1, 2)$, and Z_i $(i = 1, 2)$ be the reported response in sub-sample i $(i = 1, 2)$. Under this model, the reported response Z_i in the i^{th} sub-sample is given by

$$Z_i = \begin{cases} Y & \text{with probability 1-W} \\ Y + S_i & \text{with probability W} \end{cases} \quad where \quad i = 1, 2. \tag{I.14}$$

The expected value and variance of Z_i are given by

$$E(Z_i) = \mu_y + \mu_{s_i} W \tag{I.15}$$

and

$$Var(Z_i) = \sigma_y^2 + \sigma_{s_i}^2 W + \mu_{s_i}^2 W(1-W), \ \ where \ \ \mu_{s_i} = E(S_i) \ \ (i = 1, 2). \qquad \text{(I.16)}$$

The unbiased estimators $\hat{\mu}_y$ and $\hat{W}$ and their corresponding variances are given by

$$\hat{\mu}_y = \frac{\mu_{s_1}\bar{z_2} - \mu_{s_2}\bar{z_1}}{\mu_{s_1} - \mu_{s_2}}, \qquad \text{(I.17)}$$

$$\hat{W} = \frac{\bar{z_1} - \bar{z_2}}{\mu_{s_1} - \mu_{s_1}}, \qquad \text{(I.18)}$$

$$Var(\hat{\mu}_y) = \frac{1}{(\mu_{s_2} - \mu_{s_1})^2}(\mu_{s_2}^2 \frac{\sigma_{z_1}^2}{n_1} + \mu_{s_1}^2 \frac{\sigma_{z_2}^2}{n_2}), \qquad \text{(I.19)}$$

and

$$Var(\hat{W}) = \frac{1}{(\mu_{s_2} - \mu_{s_1})^2}(\frac{\sigma_{z_1}^2}{n_1} + \frac{\sigma_{z_2}^2}{n_2}) \quad \mu_{s_1} \neq \mu_{s_2}. \qquad \text{(I.20)}$$

I.1.6 Diana and Perri's Linear Combination Model (2011)

The goal of a RRT model is to protect respondent privacy. Diana and Perri [08] believe that a combination of additive and multiplicative approaches can bring more confidence among the respondents about their privacy protection since two scrambling variables will be introduced to the model. Let T be a scrambling variable with mean μ_T and variance σ_T^2; and S be another scrambling variable, independent of T, and with mean μ_s and variance σ_s^2. Both T and S are independent of the study variable Y. They introduced a more general linear combination model given by

$$Z = TY + S. \qquad \text{(I.21)}$$

It is common to assume μ_T=1 and μ_s=0. Then the expected value and variance of Z are given by

$$E(Z) = \mu_y, \tag{I.22}$$

and

$$Var(Z) = \sigma_s^2(\mu_y^2 + \sigma_y^2) + \sigma_y^2 + \sigma_T^2. \tag{I.23}$$

If μ_T=1 and μ_s=0, the unbiased estimator $\hat{\mu_y}$ and its variance are given by

$$\hat{\mu_y} = (\bar{z} - \mu_s)/\mu_T = \bar{z}, \tag{I.24}$$

and

$$Var(\hat{\mu_y}) = \frac{1}{n}[\sigma_s^2(\mu_y^2 + \sigma_y^2) + \sigma_y^2 + \sigma_T^2]. \tag{I.25}$$

In this section, we have introduced several RRT models and presented estimators for the sensitive variable mean, as well as the estimator for the sensitivity level. In Section I.2, we will talk about another technique that we often use for high non-response rate.

I.2. Hansen and Hurwitz Two-phase Sampling Techniques

As we mentioned earlier, non-response is widespread in email or phone surveys. Non-response refers to individuals who are chosen for the sample and are unwilling or unable to participate in the survey. Some of them may feel no obligation to complete a survey, or they do not care about the survey itself and refuse to do it; others may not be available at the time of the survey; or a person may not feel comfortable to provide the true answer for the survey question. Such cases reduce the precision of population estimates.

Since email or phone surveys are easier, cheaper and more convenient, nowadays many researches use these two survey methods to obtain information. However, the high non-response rate become an important concern in the study. According to Fan & Yan (2010) [11] and Miller & Dillman (2011)[49], a response rate of 40-50 percent is considered excellent. In reality, it is much smaller than this. Among all the sampling methods, personal face-to-face interview is the one that reduces non-response rate the most, but the cost is considerably higher than other methods. One may wonder if we could combine the strengths of different survey methods. Hansen and Hurwitz (1946) [25] were the first to suggest a procedure of taking a sub-sample of non-respondents after the first mail or phone attempt and then obtain information from the sub-sample by personal interview. The details are provided below.

Let $U = \{U_1, U_2, ..., U_N\}$ be a finite population of size N and a random sample without replacement of size n is taken. We assume that n_1 units provided response on the first call and therefore $n_2 = n - n_1$ units did not respond. Then a sub-sample of size $n_s = \frac{n_2}{f}$ $(f > 1)$ is taken from the n_2 non-response units. Hansen and Hurwitz (1946) used mail survey at the first attempt and then used face-to-face interview at the second attempt. Let $\mu_y = \frac{\sum_{i=1}^{N} y_i}{N}$ and $\sigma_y^2 = \frac{\sum_{i=1}^{N}(y_i-\mu_y)^2}{N-1}$ respectively be the population mean and variance of the study variable y. Let $\mu_{y_1} = \frac{\sum_{i=1}^{N_1} y_i}{N_1}$ and $\sigma_{y_1}^2 = \frac{\sum_{i=1}^{N_1}(y_i-\mu_{y_1})^2}{N_1-1}$ respectively be the mean and variance of response group of size N_1, and $\mu_{y_2} = \frac{\sum_{i=1}^{N_2} y_i}{N_2}$ and $\sigma_{y_2}^2 = \frac{\sum_{i=1}^{N_2}(y_i-\mu_{y_2})^2}{N_2-1}$ respectively be the mean and variance of non-response group of size N_2. Then the population mean is given by

$$\mu_y = W_1\mu_{y_1} + W_2\mu_{y_2}. \tag{I.26}$$

where $W_1 = \frac{N_1}{N}$ and $W_2 = \frac{N_2}{N}$. Not knowing N_1 poses a challenge of its own.

Let $\bar{y}_1 = \frac{\sum_{i=1}^{n_1} y_i}{n_1}$ be the sample mean for the response group, and $\bar{y}_2 = \frac{\sum_{i=1}^{n_s} y_i}{n_s}$ be the sample mean for the non-response group. One can note here that $\bar{y}_1$ and $\bar{y}_2$ are unbiased estimators for μ_{y_1} and μ_{y_2}, respectively. But $\bar{y}_1$ has a bias $W_2(\mu_{y_1} - \mu_{y_2})$ in estimating the population mean of μ_y.

Hansen and Hurwitz (1946) suggested an unbiased population mean estimator given by

$$\hat{\mu}_y = w_1\bar{y}_1 + w_2\bar{y}_2. \tag{I.27}$$

where $w_1 = \frac{n_1}{n}$ and $w_2 = \frac{n_2}{n}$. The variance of $\bar{y}$ is given by

$$Var(\hat{\mu}_y) = (\frac{N-n}{Nn})\sigma_y^2 + \frac{W_2(f-1)}{n}\sigma_{y2}^2 \tag{I.28}$$

Their results showed that the mean estimation is more efficient and accurate since we obtain more information from the population.

So far we have talked about RRT models and Hansen and Hurwitz (1946) two-phase technique which are used to reduce the social desirability bias and the participation bias, respectively. There is another error called measurement error that we mentioned earlier. We will briefly introduce it in Section I.3 below.

I.3. Measurement Errors

Measurement error is also called observational error which is the difference between observed value and the true value of a variable. It usually can be divided into two components - random error and systematic error. Random errors occur because of random and inherently unpredictable events in the measurement process. Systematic errors are errors that are not determined by chance but are a consequence of a problem in the measurement system that affects all measurements in the same way. For example, scientists study global warming and need to measure the temperatures. If the temperatures were measured with a simple thermometers and the data were recorded by hand, it may cause some random errors because people sometimes make errors in reading the thermometers or recording the temperatures. However, if the scientists' measurements are mostly temperatures near an urban area, it may cause some systematic error because all the temperatures are probably higher than in the urban area as compared to rural areas. Urban areas tend to be warmer than rural areas because of heat released by human activities. Measurement errors are very common in sampling surveys. We could reduce measurement errors by double checking all the measurements for accuracy, taking average of multiple measurements, and making sure the instrument has the highest precision etc.

Most of the time we assume measurement errors are very small and neglect them. But if measurement errors are not small enough, then we get unreliable estimates. Therefore, we will incorporate measurement errors in our mean estimation.

As mentioned at beginning, a proper sampling method can determine whether or not a sample is truly representative sample or not. In section I.4, two sampling methods will be presented to reflect on this problem.

I.4. Simple Random Sampling and Stratified Random Sampling

Simple random sampling is a sampling method where every sample of the same size has an equal change of being selected. For example, to choose a simple random sample of 10 universities from all the universities in a state, you could assign a number to each university and select a sample by letting a computer randomly generate 10 numbers. This is the most commonly used method because it is likely to provide a representative sample as long as the sample size is large enough. However, when sub-groups within a population vary, simple random sampling may not be a good choice.

Stratified random sampling is a sampling method that is used when researchers are either trying to draw conclusions from different sub-groups or strata that share some common characteristics, or when the population is not very homogeneous. While using stratified random sampling, the population is divided into different strata based on their common characteristics, such as gender, education level, geographic location, nationality and age etc. Then researchers can randomly select a simple random sample from each stratum and the estimates are aggregated over the strata.

Suppose we want to estimate the average income of individuals in a town. Assume that the town has 2000 residents with Master's degree or above; 3000 residents with college degree; and 5000 residents with high school degree or below. We may choose a simple random sample of size 100 from this town. However, because the incomes are extremely different for people with different education levels, we may get a better estimation if we collect income data from residents with each education level. We could take a proportional random sample of sizes 20, 30, and 50 from education level high to low groups, respectively. In this way, we can create a more representative sample.

Simple random sampling and stratified random sampling are two of the most important and commonly used sampling methods. Therefore, we will study mean estimation under both methods in this dissertation.

I.5. Outline of the Dissertation

Chapter I provided the background of this study and an introduction to the techniques that will be used in the study, including several RRT models, Hansen and Hurwitz (1946) two-phase sampling, and the basic idea about measurement errors.

Chapter II presents the literature review. It includes mean estimations under RRT models, measurement errors, non-response, and stratified random sampling.

Chapter III presents mean estimators under measurement errors using the ORRT model. In this Chapter, the efficiency and the privacy of a general linear combination RRT model and a simple additive RRT model will be compared. A better criterion factoring in both efficiency and privacy of a RRT model is used for the entire study. A simulation study is also conducted to show the performance of various mean estimators.

Chapter IV presents mean estimators under simultaneous presence of measurement errors and non-response using ORRT model. We will introduce a modified version of Hansen and Hurwize two-phase sampling, then study some mean estimators using this new technique. A simulation study is also conducted to show the performance of the mean estimators.

Chapter V presents mean estimators under measurement errors and non-response under stratified random sampling using ORRT model. A simulation study is also conducted to show the performance of the mean estimators under stratified sampling.

Chapter VI presents a general discussion on the research carried out in the dissertation and some future directions.

CHAPTER II

LITERATURE REVIEW

In Chapter I, we discussed the background and the techniques that will be used in the study. Some literature review will be presented in Chapter II. All the literature review revolves around the topic of mean estimation, but under different conditions. We divided this chapter into four parts - mean estimation under RRT models; RRT models and measurement errors; non-response; and stratified random sampling.

II.1. Mean Estimation under RRT Models

Researchers have been working on the mean estimation for sensitive variables for years. They have discussed different estimators under the same RRT model or the same estimator under different RRT models. Many researchers have studied the mean estimation when the primary variable is sensitive and there is no auxiliary variables. These include Gupta and Shabbir (2004)[18], Gupta et al. (2002, 2010)[17][19], Wu et al. (2008)[86], Saha (2008)[56], and Perri(2008)[52] etc. Also, many others have used auxiliary information to improve the efficiency of the estimators, such as Kadilar and Cingi (2005, 2006)[28][29], Kadilar et al. (2007)[30], Shabbir & Gupta (2007, 2010)[59][60], Turgut and Cingi(2008)[83], Nangsue(2009)[51], Koyuncu and Kadilar (2009)[44], Sousa et al. (2010)[77], Subramani and Kumarapandiyan (2012)[81], Gupta et al. (2012, 2015, 2016) [20][22][23], Tarray & Singh (2015) [82], Kalucha et al. (2015)[31], and Zhang et al. (2018)[89] etc.

In this dissertation, we focus on estimating the mean of a sensitive variable using non-sensitive auxiliary variable that is highly correlated with the primary variable.

In this section, we will discuss in detail some existing mean estimators using RRT or ORRT models.

II.1.1 Mean Estimation under RRT Models (Pollock & Bek 1976 and Sousa et al. 2010)

Ratio and product estimators provide more accurate estimates than the ordinary mean estimator when an auxiliary variable exists that is highly correlated with the study variable. In sample surveys, there are some situations when the variable of interest (Y) is sensitive but there is a nonsensitive auxiliary variable (X) which is highly correlated with it. For example, Y may be the number of sexual partners a woman might have had in her life and X may be her age. In such cases, one can estimate mean of Y using one of the RRT models and improve the estimator by using auxiliary information.

Sousa et al. (2010) [77] proposed a ratio estimator where the mean of Y is estimated using the Pollock & Bek (1976) RRT model and it is further improved by an auxiliary variable X. Again, let Y be the sensitive variable of interest. Let X be a non-sensitive auxiliary variable which is observed directly and also is positively correlated with Y. We assume that the mean (μ_x) and variance (σ_x^2) for X are known. Let S be a scrambling variable independent of Y and X. The respondents are asked to provide scrambled responses for Y given by Z=Y+S but report the true responses for X. We assume the population mean of X (μ_x) is known. And population mean of S, μ_s = E(S)=0. Thus, E(Z)=E(Y).

If the auxiliary variable X is ignored, then an unbiased ordinary estimator of μ_y is given by

$$\hat{\mu}_o = \bar{z} \tag{II.1}$$

and the MSE of $\hat{\mu}_0$ is given by

$$MSE(\hat{\mu}_0) = \lambda(\sigma_y^2 + \sigma_s^2), \tag{II.2}$$

where $\lambda = (N-n)/Nn$.

The proposed ratio estimator for the population mean of Y using the auxiliary variable X is given by

$$\hat{\mu}_R = \bar{z}\frac{\mu_x}{\bar{x}}. \tag{II.3}$$

The MSE of the estimator $\hat{\mu}_R$, correct up to first order of approximation, is given by

$$MSE^{(1)}(\hat{\mu}_R) \cong \lambda\mu_y^2(C_z^2 + C_x^2 - 2\rho_{zx}C_zC_x), \tag{II.4}$$

and correct up to second order of approximation, is given by

$$MSE^{(2)}(\hat{\mu}_R) \cong MSE^{(1)}(\hat{\mu}_R) + 3\mu_y^2\lambda^2C_x^2[(1+2\rho_{zx}^2)C_z^2 + 3C_x^2 - 6\rho_{zx}C_zC_x], \tag{II.5}$$

where $\lambda = \frac{N-n}{Nn}$.

Comparing the first order of approximation in (II.4) and (II.2), Sousa et al. (2010) [77] showed that the ratio estimator $\hat{\mu}_R$ is more efficient than the RRT mean estimator $\hat{\mu}_0$ when Y and X have strong positive correlation.

II.1.2 Mean Estimation under ORRT Models (Kalucha et al. 2015 and Zhang et al. 2018

Sousa et al. (2010) [77] was the first to use ratio estimators under RRT models. They estimated μ_y using a non-optional RRT model with the utilization of a non-sensitive auxiliary variable. However, Gupta et al. (2002) [17] introduced ORRT models and showed that they perform better than non-optional RRT. Based on this result, Gupta et al. (2014) [21] improved Sousa et al. (2010) [77] by using optional scrambling. Additionally, in the Gupta et al. (2010) [19], they use a split-sample approach using different scrambling variables in the two sub-samples. Kalucha et al. (2015) [31] and Zhang et al (2018) [89] also improved the Sousa et al. (2010) estimator further by using a split sample ORRT model.

If a proportion W of the respondents feel the survey question is sensitive, then according to Gupta et al.(2010), the reported response Z_i in the i^{th} sub-sample is given by

$$Z_i = \begin{cases} Y & \text{with probability 1-W} \\ Y + S_i & \text{with probability W} \end{cases} \qquad i = 1, 2. \tag{II.6}$$

Kalucha et al. (2015) [31] proposed two ratio estimators of finite population mean using ORRT model and called them the additive ratio estimator and the multiplicative ratio estimator, respectively. Let $\bar{x}_i$ and $\bar{z}_i$ (i=1, 2) respectively be the means of the auxiliary variable and the reported response in the i^{th} sub-sample. These estimators with associated MSEs, correct up to the first order of approximation, are given by:

$$\hat{\mu}_{AR} = (\frac{\mu_{s_2}\bar{z}_1 - \mu_{s_1}\bar{z}_2}{\mu_{s_2} - \mu_{s_1}})(\frac{\mu_x}{\bar{x}_1} + \frac{\mu_x}{\bar{x}_2})(\frac{1}{2}), \tag{II.7}$$

$$\hat{\mu}_{MR} = (\frac{\mu_{s_2}\bar{z}_1 - \mu_{s_1}\bar{z}_2}{\mu_{s_2} - \mu_{s_1}}))(\frac{\mu_x}{\bar{x}_1})(\frac{\mu_x}{\bar{x}_2}), \tag{II.8}$$

$$\begin{aligned} MSE^{(1)}(\hat{\mu}_{AR}) =& E(\mu_{AR} - \mu_y)^2 \\ \approx& \lambda_1[(\frac{\mu_{s_2}}{\mu_{s_2} - \mu_{s_1}})^2\sigma_{z_1}^2 + \frac{1}{4}\mu_y^2 C_x^2 - \mu_y \rho_{yx}\sigma_y(\frac{\mu_{s_2}}{\mu_{s_2} - \mu_{s_1}})C_x] \\ &+ \lambda_2[(\frac{\mu_{s_1}}{\mu_{s_2} - \mu_{s_1}})^2\sigma_{z_2}^2 + \frac{1}{4}\mu_y^2 C_x^2 + \mu_y \rho_{yx}\sigma_y(\frac{\mu_{s_1}}{\mu_{s_2} - \mu_{s_1}})C_x], \end{aligned} \tag{II.9}$$

and

$$\begin{aligned} MSE^{(1)}(\hat{\mu}_{MR}) =& E(\mu_{MR} - \mu_y)^2 \\ \approx& \lambda_1[(\frac{\mu_{s_2}}{\mu_{s_2} - \mu_{s_1}})^2\sigma_{z_1}^2 + \mu_y^2 C_x^2 - 2\mu_y \rho_{yx}\sigma_y(\frac{\mu_{s_2}}{\mu_{s_2} - \mu_{s_1}})C_x] \\ &+ \lambda_2[(\frac{\mu_{s_1}}{\mu_{s_2} - \mu_{s_1}})^2\sigma_{z_2}^2 + \mu_y^2 C_x^2 + 2\mu_y \rho_{yx}\sigma_y(\frac{\mu_{s_1}}{\mu_{s_2} - \mu_{s_1}})C_x]. \end{aligned} \tag{II.10}$$

Kalucha et al. (2015) showed that the additive ratio estimator $\hat{\mu}_{AR}$ is more efficient than the ordinary RRT estimator $\hat{\mu}_y$ when the correlation between the study variable and the auxiliary variable is greater than $\frac{1}{2}$. However, the multiplicative ratio estimator was not found to be as efficient as the ordinary RRT estimator ($\hat{\mu}_y$) or the additive ratio estimator ($\hat{\mu}_{AR}$). But Zhang et al (2018) [89] modified the multiplicative ratio estimator in (II.8) and proposed a new geometric mean ratio estimator. It is given by

$$\hat{\mu}_{GMR} = (\frac{\mu_{s_2}\bar{z}_1 - \mu_{s_1}\bar{z}_2}{\mu_{s_2} - \mu_{s_1}})\sqrt{(\frac{\mu_x}{\bar{x}_1})(\frac{\mu_x}{\bar{x}_2})}. \tag{II.11}$$

The MSE of $\hat{\mu}_{GMR}$, correct up to the first order of approximation, is given by

$$\begin{aligned} MSE^{(1)}(\hat{\mu}_{GMR}) \approx& \lambda_1[(\frac{\mu_{s_2}}{\mu_{s_2} - \mu_{s_1}})^2\sigma_{z_1}^2 + \frac{1}{4}\mu_y^2 C_x^2 - \mu_y \rho_{yx}\sigma_y(\frac{\mu_{s_2}}{\mu_{s_2} - \mu_{s_1}})C_x] \\ &+ \lambda_2[(\frac{\mu_{s_1}}{\mu_{s_2} - \mu_{s_1}})^2\sigma_{z_2}^2 + \frac{1}{4}\mu_y^2 C_x^2 + \mu_y \rho_{yx}\sigma_y(\frac{\mu_{s_1}}{\mu_{s_2} - \mu_{s_1}})C_x]. \end{aligned} \tag{II.12}$$

.

Comparing the MSE of the geometric mean ratio estimator with Kalucha et al. ratio estimators and the ordinary mean estimator in both the equal and the unequal sample split, they concluded that up to the first order approximation:

- The geometric mean ratio estimator is always more efficient than the multiplicative ratio estimator.
- The geometric mean ratio estimator is more efficient than the ordinary RRT estimator when the correlation coefficient between X and Y is greater than $\frac{1}{2}$.
- The geometric mean ratio estimator is as efficient as the additive ratio estimator up to the first order of approximation.

Since the MSE of the geometric mean ratio estimator, up to the first order of approximation, is same as that of the additive ratio estimator, the biases of these two estimators are compared. One can verify that

$$Bias^{(1)}(\hat{\mu}_{GMR}) - Bias^{(1)}(\hat{\mu}_{AR}) = -\frac{1}{8}\mu_y C_x^2(\lambda_1 + \lambda_2), \tag{II.13}$$

which means the geometric mean ratio estimator has less bias if the μ_y is positive.

II.2. Mean Estimation under RRT Models and Measurement Errors

As in Section II.1, there are lots of population mean estimators that have proposed under RRT models. In addition to social desirability response bias, there are some other non-sampling errors such as measurement errors that may also affect the population mean estimation. Many researches have studied measurement errors while utilizing auxiliary information, including Shalabh (1997)[62], Manisha and Singh (2001) [48], Srivastava and Shalabh (2001) [80], Allen et al. (2003) [02], Singh and

Karpe (2007, 2008, 2009, 2010)[65][66][68][70], Gregoire and Salas (2009) [16], Salas and Gregorie (2010) [55], Kumar et al. (2011) [45], Kumar et al. (2011) [46], Shukla et al. (2012) [63], Singh, V. K., Singh, R. and Smarandache (2014) [75] etc.

Although Blattman et al (2014) [04] developed a survey validation technique for qualitative variables to check for measurement errors when dealing with sensitive attributes, not much effort has been devoted to estimating the finite population mean of a sensitive variable in the presence of measurement errors. We know that use of RRT models reduces non-sampling errors when the variable of interest is sensitive, one may also want to check the impact of measurement errors. Khalil et al. (2018) [34] studied mean estimation for sensitive variables in the presence of measurement errors under a non-optional RRT model. The details are provided below.

II.2.1 Mean Estimation under RRT Models in the Presence of Measurement Errors (Khalil et al. 2018)

According to Pollock & Bek (1976) RRT model, the respondent is asked to provide a scrambled value for Y given by Z=Y+S, and report a true response for the non-sensitive auxiliary variable X. Let the measurement errors for the scrambled response variable (Z) and the auxiliary variable (X) on i_{th} unit respectively be U_i and V_i. U_i and V_i are assumed to be random and independent with mean zero and variance σ_u^2 and σ_v^2 respectively.

There are some commonly used existing mean estimators and their MSEs in the presence of measurement errors:

- The ordinary RRT mean estimator is given by

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} z_i}{n} = \bar{z}. \tag{II.14}$$

The MSE of $\hat{\mu}_0$ is given by

$$MSE^*(\hat{\mu}_0) = \lambda(\sigma_z^2 + \sigma_u^2), \tag{II.15}$$

where $\lambda = (N-n)/Nn$.

- A ratio estimator proposed by Sousa et al. (2010) is given by

$$\hat{\mu}_R = \bar{z}\frac{\mu_x}{\bar{x}}. \tag{II.16}$$

The MSE of $\hat{\mu}_R$ is given by

$$MSE^*(\hat{\mu}_R) = \lambda(\sigma_z^2 + R^2\sigma_x^2 - 2R\rho_{zx}\sigma_x\sigma_z) + \lambda(\sigma_u^2 + R^2\sigma_v^2), \tag{II.17}$$

where $R = \mu_z/\mu_x$.

Khalil et al. (2018) generalized the estimator in (II.16) and proposed a generalized randomized response estimator for the mean of a sensitive study variable Y in the presence of a highly correlated (positively) auxiliary variable and measurement errors. The proposed estimator is given by

$$\hat{\mu}_N = (\bar{z} + k(\mu_x - \bar{x}))(\frac{\bar{W}}{\bar{w}})^g, \tag{II.18}$$

where $\bar{w} = \phi(\alpha\bar{x}+\beta)+(1-\phi)(\alpha\mu_x+\beta)$, $\bar{W} = \alpha\mu_x+\beta$, k and g are suitable constants, and ϕ is assumed to be an unknown constant whose value is to be determined from optimality considerations. α ($\alpha \neq 0$) and β are assumed to be some known parameters of the auxiliary variable X, such as coefficient of variation (C_x), kurtosis, and correlation coefficient (ρ_{zx}) etc.

The minimum MSE of $\hat{\mu}_N$, correct up to the first order of approximation, is given by

$$MSE *_{min} (\hat{\mu}_N) \approx \lambda(\sigma_z^2 + \sigma_u^2 - \frac{\rho_{zx}^2 \sigma_z^2 \sigma_x^2}{\sigma_x^2 + \sigma_v^2}). \tag{II.19}$$

Their result showed that MSE increases when measurement errors exist. And the generalized estimator ($\hat{\mu}_N$) is more efficient than the ordinary RRT mean estimator ($\hat{\mu}_o$) and ratio estimator ($\hat{\mu}_R$) both with and without measurement errors, particularly if Y and X are highly correlated.

Khalil et al. in their study used what are known as full RRT (or non-optional RRT models) where all respondents provide a scrambled response. As mentioned earlier, Gupta et al. (2002) [17] ORRT model is generally more efficient than the corresponding non-optional RRT model. Also, in a recent publication by Gupta et el. (2018)[24], it is shown that there is no extra loss of privacy in using ORRT models as compared to the corresponding RRT models. So why not use ORRT models instead of full RRT model? Using ORRT model in this situation has not been done by anyone so far. This is our main motivation for this dissertation. The work will be introduced in Chapter III.

II.3. Mean Estimation under Non-response

Non-response is another common non-sampling error we have seen in sampling. The problem of non-response has been discussed in many papers. These include Hansen and Hurwitz (1946) [25], Foradori (1961) [12], Srinath (1971) [79], Khare and Srivastava (1993, 1995, 1997, and 2010) [37][38][39][40], Singh and Kumar (2008, 2009, and 2011) [67][69][73], Singh et al. (2010) [71], Kumar and Bhougal (2011) [47], Shabbir and Khan (2013) [61], Singh and Sharma (2015)[76], and Azzem and Hanif

(2017) [03] etc. Most of these researchers suggested different types of estimators for population parameters based on Hansen and Hurwitz (1946) double sampling plan. Some of them also used different conditions such as mean estimation in the presence of both non-response and measurement errors, or mean estimation under non-response and RRT models.

II.3.1 Mean Estimation under Non-response and Measurement Errors (Singh and Sharma 2015)

Singh and Sharma (2015)[76] have studied the problem of estimating the finite population mean in the presence of non-response and measurement errors. In their study, they assume non-response and measurement errors happened on both the study and auxiliary variables and utilized Hansen and Hurwitz (1946) two-phase sampling. Assume that n_1 units provided response on the first call and $n_2 = n - n_1$ units did not respond. Then a sub-sample of size $n_s = \frac{n_2}{f}$ $(f > 1)$ is taken from the n_2 non-responding units. Let N_1 and N_2 respectively be the sizes of the respondent group and the non-respondent group in the population.09 The proportions of the response group and the non-response group in the population are $W_1 = \frac{N_1}{N}$ and $W_2 = \frac{N_2}{N}$, respectively. Let the first phase measurement errors on the study variable Y and auxiliary variable X on the i^{th} unit be U_i and V_i; and let the second phase measurement errors of Y and X on the i^{th} unit be U_{2i} and V_{2i}. Assume these measurement errors are random and independent with variances of σ_u^2, σ_v^2, σ_{u2}^2 and σ_{v2}^2, respectively. In order to compare efficiency of mean estimators, some adapted estimators under both Hansen and Hurwitz (1946) two-phase sampling and measurement errors are here.

- The ordinary mean estimator is given by

$$\hat{\mu}_0 = w_1\bar{y_1} + w_2\bar{y_2}, \tag{II.20}$$

where $w_1 = \frac{n_1}{n}$ and $w_2 = \frac{n_2}{n}$.The expected value of $\hat{\mu}_0$ is given by

$$E(\hat{\mu}_0) = W_1\mu_{y_1} + W_2\mu_{y_2} = \mu_y. \tag{II.21}$$

The variance of $\hat{\mu}_0$ under measurement errors is given by

$$Var(\hat{\mu}_0) = \theta(\sigma_y^2 + \sigma_u^2) + \lambda(\sigma_{y_{(2)}}^2 + {\sigma_{u}}_{(2)}^2), \tag{II.22}$$

where $\theta = \frac{N-n}{Nn}$ and $\lambda = \frac{W_2(f-1)}{n}$.

- A ratio estimator is given by

$$\hat{\mu}_R = \frac{\bar{y}^*}{\bar{x}^*}\mu_x. \tag{II.23}$$

The MSE of $\hat{\mu}_R$ under measurement errors is given by

$$\begin{aligned} MSE(\hat{\mu}_R) =& \theta\mu_y^2(C_y^2 + C_x^2 + \frac{\sigma_u^2}{\mu_y^2} + \frac{\sigma_v^2}{\mu_x^2} - 2\rho_{yx}C_yC_x) + \\ & \lambda\mu_y^2(\sigma_{y_{(2)}}^2 + \sigma_{x_{(2)}}^2 + \frac{{\sigma_{u}}_{(2)}^2}{\mu_y^2} + \frac{{\sigma_{v}}_{(2)}^2}{\mu_x^2}). \end{aligned} \tag{II.24}$$

Singh and Sharma (2015) proposed a class of estimators given by

$$\hat{\mu}_P = m_1\bar{y}^* + m_2\frac{\bar{y}^*}{\bar{x}^*}\mu_x. \tag{II.25}$$

Their proposed class of estimators is a specific version of the ordinary estimator and ratio estimator if we let $(m_1, m_2) = (1, 0)$ and $(m_1, m_2) = (0, 1)$, respectively. By taking the optimum values of (m_1, m_2) with $m_2^* = \frac{1}{R}[\frac{O}{M}]$ and $m_1^* = 1$ - m_2^*, where R

$= \frac{\mu_y}{\mu_x}$, the minimum MSE of the mean estimator $\hat{\mu}_P$ is given by

$$MSE(\hat{\mu}_P) = M[1 - \frac{O^2}{MN}] \tag{II.26}$$

where M $= \frac{1}{n}(\sigma_y^2 + \sigma_u^2) + \frac{(k-1)W_2}{n}(\sigma_{y(2)}^2 + \sigma_{u(2)}^2)$, N $= \frac{1}{n}(\sigma_x^2 + \sigma_v^2) + \frac{(k-1)W_2}{n}(\sigma_{x(2)}^2 + \sigma_{v(2)}^2)$ and O $= \frac{1}{n}\rho_{yx}\sigma_x\sigma_y + \frac{(k-1)W_2}{n}\rho_{xy_2}\sigma_{x(2)}\sigma_{y(2)}$. Their results shows that $\hat{\mu}_P$ is more efficient than $\hat{\mu}_0$ if

$$M - M(1 - \frac{O^2}{MN}) = \frac{O^2}{MN} > 0; \tag{II.27}$$

and $\hat{\mu}_P$ is more efficient than $\hat{\mu}_R$ if

$$(M + N - 2O) - M(1 - \frac{O^2}{MN}) = N - 2O + \frac{O^2}{MN} > 0. \tag{II.28}$$

II.3.2 Mean Estimation under Non-response and RRT Models

As we mentioned in the previous sections, the respondents are unlikely to provide true response in face-to-face interview if the survey question is sensitive. To reduce the bias caused by sensitive questions, one could use randomized response technique (RRT) models when we target the non-response group. Respondents may refuse to respond on the first call but may provide scrambled response on the second call with personal interview. Diana et al. (2014)[09] proposed an unbiased population mean estimator under Hansen and Hurwitz (1946) two-phase sampling. Their estimator reduces non-response but increases the estimator variance due to the use of RRT model in the non-respondent group. Later, Ahmed et al. (2017) [01] proposed generalized ratio and regression estimators utilizing known coefficient of variation of the study variable in case of second sample by using RRT approach. This estimator improved the efficiency when the auxiliary variable and the study variable are highly correlated. Here we only discuss Diana et al. (2014) in detail.

Diana et al. (2014) used a RRT model in the second phase non-respondents group where the scrambled response is given by Z=TY+S. T and S are two scrambling variables that are independent of Y. Then a modified version of Hansen and Hurwitz (1946) estimator is given by

$$\hat{\mu}_{0HH} = w_1\bar{y}_1 + w_2\hat{\bar{y}}_2, \tag{II.29}$$

where $\hat{\bar{y}}_2 = \sum_{i=1}^{n_s}(\frac{z_i}{n_s})$ and z_i is the scrambled response from the second face-to-face interview step. The variance of the unbiased estimator $\hat{\bar{y}}^*$ is given by

$$Var(\hat{\mu}_{0HH}) = \theta\sigma_y^2 + \lambda\sigma_{y_{(2)}}^2 + G, \tag{II.30}$$

where G $= \frac{W_2 h}{n}[\frac{\sigma_T^2(\sigma_{y(2)}+\mu_{y_{(2)}}^2)+\sigma_s^2+2\sigma_{st}\mu_y^2}{\mu_T^2}]$ is the penalty for using RRT models.

They suggested a regression estimator where auxiliary information is used. Assume the population mean μ_x of the auxiliary variable is known and non-response only happened on Y. The estimator is given by

$$\hat{\mu}_{regHH} = \hat{\mu}_{0HH} + \hat{\beta}_{yx}^*(\mu_x - \bar{x}), \tag{II.31}$$

where $\hat{\mu}_{0HH}$ is the modified version of the Hansen and Hurwitz (1946) estimator and $\hat{\beta}_{yx}^* = \frac{\hat{\sigma}_{yx}^*}{\sigma_x^2}$. The MSE of $\hat{\mu}_{regHH}$ is given by

$$MSE(\hat{\mu}_{regHH}) = \theta\sigma_y^2(1-\rho_{yx}^2) + \lambda\sigma_{y_{(2)}}^2 + G. \tag{II.32}$$

Diana et al. (2014) also considered a situation when non-response is present in the auxiliary variable X, and also suggested a regression estimator

$$\hat{\mu}_{regHH1} = \hat{\mu}_{0HH} + \hat{\beta}_{yx}^{**}(\mu_x - \bar{x}^*), \tag{II.33}$$

where $\hat{\beta}^{**}_{yx} = \frac{\hat{\sigma}^*_{yx}}{\sigma^{*2}_x}$. The MSE of $\hat{\mu}_{regHH1}$ is given by

$$MSE(\hat{\mu}_{regHH1}) = \theta\sigma^2_y(1-\rho^2_{yx}) + \lambda(\sigma^2_{y_{(2)}} + \beta^2_{yx}\sigma_{x(2)} - 2\beta_{yx}\sigma_{yx_{(x)}}) + \lambda\sigma^2_{y_{(2)}} + G. \tag{II.34}$$

Researchers have studied mean estimation under non-response alone, both non-response and measurement errors, and both non-response and RRT models. But not many researchers have explored the performance of mean estimators for a sensitive variable under both non-response and measurement errors using ORRT model. Adding non-response to mean estimation for a sensitive question in the presence of measurement errors will be our second major motivation for this study. The work will be introduced in Chapter IV.

II.4. Mean Estimation under Stratified Random Sampling

In addition to simple random sampling, stratified random sampling is another commonly used method when subpopulations within an overall population vary. Much work has been done when study variables are directly observed in stratified random sampling, including Kadilar and Cingi (2003, 2005)[26][27], Shabbir and Gupta (2005,2006) [57][58], Koyuncu and Kadilar (2008, 2009, 2010)[41][42][43], Singh and Karpe (2010)[72], Zahid and Shabbir (2018)[88], and Khalil et al. (2017) [33] etc.. Some of this work also involves non-response, measurement errors, and sensitive questions.

II.4.1 Mean Estimation under Stratified Random Sampling in the Presence of Measurement Errors and Non-response (Zahid and Shabbir 2018)

Zahid and Shabbir (2018)[88] proposed a class of estimators in the presence of measurement errors and non-response under stratified random sampling. Assume measurement errors are found in both the study variable Y and the auxiliary variable X and non-response happened in each stratum. Their study used Hansen and Hurwitz (1946) two-phase sampling to reduce the impact of non-sampling errors caused by non-response in each stratum. Let a finite population U $= (U_1, U_2, U_3, .., U_N)$ be divided into L homogeneous strata, and N_h represent the number of units in stratum h such that $\sum_{h=1}^{L} N_h = N$. Let X and Y have population means $\mu_{xh} = \frac{1}{N_h}\sum_{i=1}^{N_h} x_{hi}$ and $\mu_{yh} = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$ respectively in stratum h. Let the respective measurement errors on the study variable Y and the auxiliary variable X in the h^{th} stratum be given by U_{hi} and V_{hi}. These measurement errors are assumed to be uncorrelated and having normal distribution with zero mean and variance σ^2_{uh} and σ^2_{vh}, respectively. It is also assumed that the measurement errors are independent of Y and X. Under Hansen and Hurwize two-phases sampling, n_{1h} units provided response on the first call and remaining $n_{2h} = n_h - n_{1h}$ units do not respond. Then a sub-sample of size $n_{sh} = \frac{n_{2h}}{f_h}$ $(f_h > 1)$ is taken from the n_{2h} non-response units in the h^{th} stratum.

The study provided some existing estimators under stratified random sampling and measurement errors as discussed below.

- The ordinary Hansen and Hurwitz (1946) mean estimator is given by

$$\hat{\mu}^*_o = \sum_{h=1}^{L} W_h \bar{y}^*_h, \tag{II.35}$$

where $\bar{y}^*_h = (\frac{n_{1h}}{n_h})\bar{y}_{1h} + (\frac{n_{2h}}{n_h})\bar{y}_{2h}$ and $W_h = N_h/N$. The expected value of $\hat{\mu}^*_o$ is given by

$$E(\hat{\mu}_o^*) = \sum_{h=1}^{L} W_h \mu_{y_h} = \mu_y. \tag{II.36}$$

The variance of $\hat{\mu}_o^*$ is given by

$$Var(\hat{\mu}_o^*) = \sum_{h=1}^{L} W_h^2 [\theta_h(\sigma_{yh}^2 + \sigma_{uh}^2) + \lambda_h(\sigma_{y(2)h}^2 + \sigma_{u(2)h}^2)], \tag{II.37}$$

where $\theta_h = \frac{N_h - n_h}{N_h n_h}$, $\lambda_h = \frac{N_{2h}(f_h - 1)}{N_h n_h}$.

- The ratio estimator is given by

$$\hat{\mu}_r^* = \sum_{h=1}^{L} W_h \frac{\bar{y}_h^*}{\bar{x}_h^*} \mu_{xh}. \tag{II.38}$$

The MSE of $\hat{\mu}_r^*$ is given by

$$\begin{aligned} MSE(\hat{\mu}_r^*) = &\sum_{h=1}^{L} W_h^2 [\theta_h(\sigma_{yh}^2 + \sigma_{uh}^2) + \lambda_h(\sigma_{y(2)h}^2 + \sigma_{u(2)h}^2) + \\ &R_h^2(\theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma_{x(2)h}^2 + \sigma_{v(2)h}^2)) - \\ &2R_h(\theta_h \rho_{yxh} \sigma_{yh} \sigma xh + \lambda_h \rho_{yx_{(2)}})], \end{aligned} \tag{II.39}$$

where $R_h = \frac{\mu_{yh}}{\mu_{xh}}$.

Zahid and Shabbir (2018) proposed a class of estimators given by

$$\hat{\mu}_{ZS}^* = \sum_{h=1}^{L} W_h [m_{1h} \bar{y}_h^* + m_{2h}(\mu_{xh} - \bar{x}_h^*)(\frac{\mu_{xh}}{\bar{x}_h^*})^{a_h} exp(1 - \alpha_h)(\frac{\mu_{xh} - \bar{x}_h^*}{\mu_{xh} + \bar{x}_h^*})], \tag{II.40}$$

where m_{1h} and m_{2h} are constants whose values are to be determined and α_h is a scalar.

By substituting optimal values of m_{1h} and m_{2h}, the minimum MSE of $\hat{\mu}^*_{ZS}$ is given by

$$MSE(\hat{\mu}^*_{ZS}) = \sum_{h=1}^{L} W_h^2[\mu_{yh}^2 - \frac{A_{h1}E_{h1}^2 + B_{h1}D_{h1}^2 - 2C_{h1}D_{h1}E_{h1}}{A_{h1}B_{h1} - C_{h1}^2}], \tag{II.41}$$

where $A_{h1} = \mu_{yh}^2 + A_h + e_h^2 t_h^2 R_h C_h + 4e_h t_h R_h^2 C_h + 2f_h t_h^2 R_h^2 B_h$,

$B_{h1} = t_h^2 B_h$, $C_{h1} = t_h C_h + 2e_h t_h^2 R_h B_h$,

$D_{h1} = \mu_{yh}^2 + e_h t_h R_h^2 C_h + f_h t_h^2 R_h^2 B_h$,

$E_{h1} = e_h t_h^2 R_h B_h$,

$A_h = \theta_h(\sigma_{yh}^2 + \sigma_{uh}^2) + \lambda_h(\sigma_{y(2)h}^2 + \sigma_{u(2)h}^2)$,

$B_h = \theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma_{x(2)h}^2 + \sigma_{v(2)h}^2)$,

and $C_h = \theta_h \rho_{yxh}\sigma_{yh}\sigma_{xh} + \lambda_h \rho_{yx_{(2)}}$.

The proposed estimator $\hat{\mu}^*_{ZS}$ is more efficient than the ordinary and ratio estimators when the following conditions hold:

- $MSE(\hat{\mu}^*_{ZS}) < Var(\hat{\mu}^*_o)$ if
  $\sum_{h=1}^{L} W_h^2 \mu_{yh}^2 - \sum_{h=1}^{L} W_h^2 \frac{A_{h1}E_{h1}^2 + B_{h1}D_{h1}^2 - 2C_{h1}D_{h1}E_{h1}}{A_{h1}B_{h1} - C_{h1}^2} - \sum_{h=1}^{L} W_h^2 A_h < 0$
- $MSE(\hat{\mu}^*_{ZS}) < MSE(\hat{\mu}^*_r)$ if
  $\sum_{h=1}^{L} W_h^2 \mu_{yh}^2 - \sum_{h=1}^{L} W_h^2 \frac{A_{h1}E_{h1}^2 + B_{h1}D_{h1}^2 - 2C_{h1}D_{h1}E_{h1}}{A_{h1}B_{h1} - C_{h1}^2} - \sum_{h=1}^{L} W_h^2 (A_h R_h^2 B_h - 2R_h C_h) < 0$

II.4.2 Mean Estimation under Stratified Random Sampling using RRT in the Presence of Measurement Errors (Khalil et al. 2018)

In Chapter II.2, we introduced Kahlil et al. (2018) [34] mean estimation of a sensitive variable in the presence of measurement errors. The study used simple random sampling. It has been further extended to stratified random sampling in

Kahlil et al (2018)[35]. In the new study, they have modified some of the existing mean estimators in the context of measurement errors under stratified random sampling:

- The ordinary mean estimator:

$$\hat{\mu}_o^{st} = \bar{z}_{st} = \sum_{h=1}^{L} W_h \bar{z}_h. \tag{II.42}$$

  The MSE of $\hat{\mu}_o^{st}$ is given by

$$MSE(\hat{\mu}_o^{st}) = \sum_{h=1}^{L} W_h^2 \theta_h (\frac{\sigma_{z_h}^2}{\gamma_{z_h}}), \tag{II.43}$$

  where θ_h=$\frac{N_h - n_h}{N_h n_h}$ and $\gamma_{z_h} = \frac{\sigma_{z_h}^2}{\sigma_{z_h}^2 + \sigma_{u_h}^2}$.

- The ratio estimator:

$$\hat{\mu}_r^{st} = \frac{\bar{z}_{st}}{\bar{x}_{st}} \mu_x = \sum_{h=1}^{L} W_h \frac{\bar{z}_h}{\bar{x}_h} \mu_x. \tag{II.44}$$

  The MSE of $\hat{\mu}_r^{st}$ is given by

$$MSE(\hat{\mu}_r^{st}) = \sum_{h=1}^{L} W_h^2 \theta_h [\frac{\sigma_{z_h}^2}{\gamma_{z_h}} + R\frac{\sigma_{x_h}}{\gamma_{x_h}}(R - 2\beta_{zx_h}\gamma_{x_h})], \tag{II.45}$$

  where $\gamma_{z_h} = \frac{\sigma_{yx_h}}{\sigma_{x_h}^2} = \frac{\sigma_{zx_h}}{\sigma_{x_h}^2}$, $\gamma_{x_h} = \frac{\sigma_{x_h}^2}{\sigma_{x_h}^2 + \sigma_{v_h}^2}$ and $R = \frac{\mu_y}{\mu_x}$.

Kahlil et al (2018)[35] proposed a generalized mean estimator for the mean of a sensitive study variable Y in the presence of measurement error, which is given by

$$\hat{\mu}_{GE}^{st} = [\bar{z}_{st} + k(\mu_x - \bar{x}_{st})][\frac{\alpha_{st}\mu_x + \beta_{st}}{w(\alpha_{st}\bar{x}_{st} + \beta_{st}) + (1-w)(\alpha_{st}\mu_x + \beta_{st})}]^g, \tag{II.46}$$

where k and g are suitable constants, and w is an unknown constant whose value is to be determined from optimality consideration. α_{st} $(\alpha_{st} \neq 0)$ and β_{st} are assumed to be some known parameters of the auxiliary variable X, such as coefficient of variation

(C_x), kurtosis, and correlation coefficient (ρ_{zx}) etc. The optimum value of $gw\phi$ which gives the minimum MSE is given by

$$(gw\phi)_{opt} = \frac{\mu_x}{\mu_y}\left(\frac{\sum_{h=1}^{L} W_h^2 \gamma_h \sigma_{zxh}}{\sum_{h=1}^{L} W_h^2 \gamma_h \sigma_{xh}^2/\theta_{xh}} - k\right). \tag{II.47}$$

By substituting optimal value of $(gw\phi)_{opt}$, the minimum value of $MSE^*(\hat{\mu}_{GE}^{st})$ is given by

$$MSE^*(\hat{\mu}_{GE}^{st}) \approx \sum_{h=1}^{L} \frac{W_h^2 \gamma_h \sigma_{zh}^2}{\theta_{zh}}(1-\rho_c^2), \tag{II.48}$$

where

$$\rho_c = \frac{\sum_{h=1}^{L} W_h^2 \gamma_h \sigma_{zxh}}{\sqrt{\sum_{h=1}^{L} W_h^2 \gamma_h \sigma_{zh}^2/\theta_{zh}}\sqrt{\sum_{h=1}^{L} W_h^2 \gamma_h \sigma_{xh}^2/\theta_{xh}}}. \tag{II.49}$$

Kahlil et al.(2018) showed that the proposed mean estimator $\hat{\mu}_{GE}^{st}$ is more efficient than the ordinary mean estimator and the ratio estimator when measurement errors are both present and absent, particularly when the study variable and the auxiliary variable are highly correlated.

Researchers have studied mean estimation under stratified random sampling and non-response; and under stratified random sampling, measurement errors, and RRT models. But not many researchers have explored the performance of mean estimators for a sensitive variable under both non-response and measurement errors using stratified random sampling. Using stratified random sampling for estimating the population mean of a sensitive variable in the presence of measurement errors and non-response simultaneously will be our third major objective for this dissertation. The work will be introduced in Chapter V.

CHAPTER III

MEAN ESTIMATION UNDER ORRT MODELS IN THE PRESENCE OF MEASUREMENT ERRORS

As mentioned in Chapter II.2, we will re-examine Khalil et al. (2018) [34] mean estimation of a sensitive variable in the presence of measurement errors but using ORRT models in this Chapter[36]. A RRT model could have different scrambling methods such as simple additive scrambling or multiplicative scrambling. In this Chapter, we also consider a broader class of scrambling methods for RRT models. In Sections III.1 and III.2, a general scrambling RRT model as well as its ORRT version will be discussed; in Section III.3, some existing mean estimators under ORRT model in the presence of measurement errors will be presented; a generalized mean estimator will be introduced in Section III.4; Section III.5 will present the simulation results; and Section III.6 will provide concluding remarks of this Chapter.

III.1. A General Scrambling Model

Let us introduce the notations again. Let Y be the sensitive study variable with unknown mean μ_y and unknown variance σ_y^2, and X be a non-sensitive auxiliary variable with known mean μ_x and known variance σ_x^2. Suppose X has a strong positive correlation with Y. Let T and S be two scrambling variables with known variances σ_T^2 and σ_s^2, respectively. Usually we choose T with a mean (μ_T) of 1 and S with a mean (μ_s) of 0. T, S, X and Y are mutually independent. Let W be the probability that the respondent finds the question sensitive. The respondent is asked to report a scrambled response for study variable (Y) if he/she considers the question sensitive, and a correct response otherwise. One could add noises to the study variable Y differently.

The most commonly used RRT model for quantitative response is the additive model given by Pollock and Bek (1976) [53] where the reported response is

$$Z = Y + S. \tag{III.1}$$

Eichhorn and Hayre (1983)[10] proposed a multiplicative model given by

$$Z = YS. \tag{III.2}$$

Diana and Perri (2011) [08] introduced a more general linear combination model given by

$$Z = TY + S. \tag{III.3}$$

As mentioned previously, it is common to assume that E(T)=μ_T=1 and E(S)=μ_s=0 in model (III.3). One can easily note that models (III.1) and (III.2) are special cases of (III.3) if we assume σ_s^2=0 and σ_T^2=0, respectively.

The multiplicative scrambling compromises respondent anonymity and it is not very efficient. Hense, in this study only the other two models will be considered. It is easy to verify that in Pollock and Bek (1976) [53] additive model (III.1), E(Z)=E(Y)=μ_y and

$$Var(Z) = \sigma_z^2 = \sigma_y^2 + \sigma_s^2; \tag{III.4}$$

and for Diana and Perri (2011) [08] general model (III.3), E(Z)=E(Y)=μ_y and

$$Var(Z) = \sigma_z^2 = \sigma_T^2(\mu_y^2 + \sigma_y^2) + \sigma_y^2 + \sigma_s^2. \tag{III.5}$$

The comparison of variances in (III.1) and (III.3) indicates that the additive model is more efficient than the general model. However, efficiency is not the only criterion that we use to evaluate RRT models. The primary objective of a RRT model is to protect respondents' privacy. Privacy level could be another consideration to evaluate RRT models.

Using the privacy protection measure $\Delta = E(Z-Y)^2$ proposed by Yan et al. (2008)[87], we can easily calculate the privacy level of the Pollock & Bek (1976) [53] model in (III.1) and the Diana & Perri (2011) [08] model in (III.3). These are given respectively by

$$\Delta_{PB} = \sigma_s^2 \tag{III.6}$$

and

$$\Delta_{DP} = \sigma_T^2(\mu_y^2 + \sigma_y^2) + \sigma_s^2. \tag{III.7}$$

One can easily notice by comparing (III.6) with (III.7) that the Diana and Perri (2011) model offers greater privacy.

Efficiency and privacy are two important considerations we use to compare RRT models. If efficiency is same, a model with higher privacy is preferred; and if privacy is same, we choose a model with better efficiency. However, neither efficiency nor privacy is need to be kept fixed here. Instead of holding one of the measures constant, Gupta et al.(2018)[24] proposed a unified measure of model quality given by

$$\delta = \frac{Var(\hat{\mu})}{PL}, \tag{III.8}$$

where $\hat{\mu}$ is the mean estimator and PL is the privacy level for the model as given by Yan et al. (2009). In (III.8), $Var(\hat{\mu})$ can be replaced by $MSE(\hat{\mu})$ in case of biased estimators. The goal of this measure is to achieve a right trade-off between efficiency and privacy protection.

One may note that the model with smaller δ value is preferred in terms of either a larger privacy level or smaller value of $Var(\hat{\mu})$. It may be observed that

$$\delta_{PB} = 1 + \frac{\sigma_y^2}{\sigma_s^2} > 1 + \frac{\sigma_y^2}{\sigma_s^2 + \sigma_T^2(\mu_y^2 + \sigma_y^2)} = \delta_{DP}. \tag{III.9}$$

Hense, while working with the general RRT model will put a burden on the model efficiency, it is better in terms of the unified measure of both efficiency and privacy. Therefore, Diana and Perri (2011) model will be used in the current study, but with a reasonably small value of σ_T^2.

III.2. ORRT Version of the General Scrambling Model

Since ORRT model is more efficient, we add optionally to the Diana and Perri (2011) model. In the ORRT version, the respondent may answer in the following two ways depending on whether the respondent considers the question sensitive or not:

$$Z = \begin{cases} Y & \text{with probability 1-W} \\ TY + S & \text{with probability W,} \end{cases} \tag{III.10}$$

where it is assumed that μ_T =E(T)=1 and μ_s =E(S)=0.

The mean and variance of scrambled response (Z) are respectively given by:

$$\begin{aligned} E(Z) &= E(Y)(1-W) + E(TY+S)W \\ &= E(Y) - E(Y)W + E(T)E(Y)W + E(S)W \\ &= E(Y) - E(Y)W + (1)E(Y)W + (0)W \\ &= E(Y) \end{aligned} \tag{III.11}$$

and

$$
\begin{aligned}
Var(Z) &= E(Z^2) - E^2(Z) \\
&= E(Y^2)(1-W) + E[(TY+S)^2]W - E^2(Z) \\
&= E(Y^2) - E(Y^2)W + E(T^2Y^2)W + 2E(TYS)W + E(S^2)W - E^2(Z) \\
&= Var(Y) + Var(S)W - E(Y^2)W + E(T^2)E(Y^2)W \\
&= Var(Y) + Var(S)W - E(Y^2)W + [Var(T)+1]E(Y^2)W \\
&= Var(Y) + Var(S)W + Var(T)E(Y^2)W \\
&= Var(Y) + Var(S)W + Var(T)[Var(Y) + E^2(Y)]W \\
&= \sigma_y^2 + \sigma_s^2 W + \sigma_T^2(\sigma_y^2 + \mu_y^2)W.
\end{aligned}
\tag{III.12}
$$

Note that $Var(Z)$ increases with W, and hence there is gain in efficiency compared to the non-optional model where W=1.

Note that under the ORRT model, the correlation coefficient between Z and X is given by

$$
\begin{aligned}
\rho_{zx} &= \frac{\sigma_{zx}}{\sqrt{\sigma_x^2}\sqrt{\sigma_z^2}} \\
&= \frac{\sigma_{zx}}{\sqrt{\sigma_x^2}\sqrt{\sigma_y^2 + \sigma_s^2 W + \sigma_T^2(\sigma_y^2+\mu_y^2)W}} \\
&= \frac{\rho_{yx}}{\sqrt{1 + \frac{\sigma_s^2 W}{\sigma_y^2} + \frac{\sigma_T^2(\sigma_y^2+\mu_y^2)W}{\sigma_y^2}}}.
\end{aligned}
\tag{III.13}
$$

This will be utilized later.

III.3. Some Existing Mean Estimators under Measurement Errors

Let a simple random sample of size n be drawn without replacement from a finite population U=$(U_1, U_2, ..., U_N)$. Let (x_i, y_i, z_i) be the observed values (factoring in measurement errors) and (X_i, Y_i, Z_i) be true values for the auxiliary variable X, the study variable Y and the scrambled response variable Z respectively associated with the i^{th} (i=1,2,...,n) sample unit. The respective measurement errors associated with the scrambled response variable (Z) and the auxiliary variable (X) are given by

$$P_i = z_i - Z_i \tag{III.14}$$

and

$$V_i = x_i - X_i. \tag{III.15}$$

These measurement errors are assumed to be random and uncorrelated with mean zero and variance σ_p^2 and σ_v^2, respectively. Some other necessary notations are given below. Let

$$\Omega_z = \sum_{i=1}^{n}(z_i - \mu_y), \tag{III.16}$$

$$\Omega_x = \sum_{i=1}^{n}(x_i - \mu_x), \tag{III.17}$$

$$\Omega_p = \sum_{i=1}^{n} P_i, \tag{III.18}$$

and

$$\Omega_v = \sum_{i=1}^{n} V_i \tag{III.19}$$

Let $e_0^* = \frac{1}{n\mu_y}(\Omega_z + \Omega_p)$ and $e_1^* = \frac{1}{n\mu_x}(\Omega_x + \Omega_v)$. In other words, $\bar{z}^* = (1 + e_0^*)\mu_y$ and $\bar{x}^* = (1 + e_1^*)\mu_x$. Under the assumption of bivariate normality (Sukhatme et al.1970)[78]:

$$E(e_0^*) = 0, \tag{III.20}$$

$$E(e_1^*) = 0, \tag{III.21}$$

$$E(e_0^{*2}) = \frac{1}{\mu_y^2}\theta(\sigma_z^2 + \sigma_p^2), \tag{III.22}$$

$$E(e_1^{*2}) = \frac{1}{\mu_x^2}\theta(\sigma_x^2 + \sigma_v^2) \tag{III.23}$$

and

$$E(e_0^* e_1^*) = \theta\rho_{zx}\frac{\sigma_z}{\mu_y}\frac{\sigma_x}{\mu_x}, \tag{III.24}$$

where $\theta = (N - n)/Nn$.

Some existing mean estimators in the presence of measurement errors using ORRT model are given below.

- The ordinary mean estimator is given by

$$\hat{\mu}_{yw} = \frac{\sum_{i=1}^{n} z_i}{n} = \bar{z}^*. \tag{III.25}$$

  It can be written as

$$\hat{\mu}_{yw} = (1 + e_0^*)\mu_y. \tag{III.26}$$

The difference between the ordinary mean estimator and the true mean can be written as

$$\hat{\mu}_{yw} - \mu_y = e_0^* \mu_y. \tag{III.27}$$

Taking square and then expected value on both side of (III.27), the MSE of $\hat{\mu}_{yw}$ is given by

$$\begin{aligned} MSE^*(\hat{\mu}_{yw}) &= \theta(\sigma_z^2 + \sigma_p^2) \\ &= \theta(\sigma_y^2 + \sigma_s^2 W + \sigma_T^2(\sigma_y^2 + \mu_y^2)W + \sigma_p^2). \end{aligned} \tag{III.28}$$

- A ratio estimator corresponding to the one in Gupta et al. (2014) is given by

$$\hat{\mu}_{rw} = \frac{\bar{z}^*}{\bar{x}^*}\mu_x = \hat{R}_w^* \mu_x. \tag{III.29}$$

It can be written as

$$\begin{aligned} \hat{\mu}_{rw} &= \frac{(1+e_0^*)\mu_y}{(1+e_1^*)\mu_x}\mu_x \\ &= \mu_y(1+e_0^*)(1+e_1^*)^{-1} \\ &= \mu_y(1+e_0^*)(1-e_1^*+e_1^{*2}-e_1^{*3}+...) \\ &= \mu_y(1-e_1^*+e_1^{*2}+e_0^*-e_0^*e_1^*+...). \end{aligned} \tag{III.30}$$

Using second order approximation, the difference between the ratio estimator and the true mean can be written as

$$\hat{\mu}_{yw} - \mu_y = \mu_y(-e_1^* + e_1^{*2} + e_0^* - e_0^* e_1^*). \tag{III.31}$$

Taking square and then expected value on both side of (III.31), the MSE of $\hat{\mu}_{rw}$ is given by

$$MSE^*(\hat{\mu}_{rw}) = \theta(\sigma_z^2 + R_w^2\sigma_x^2 - 2R_w\rho_{zx}\sigma_x\sigma_z) + \theta(\sigma_p^2 + R_w^2\sigma_v^2), \tag{III.32}$$

where $\sigma_z^2 = \sigma_y^2 + \sigma_s^2 W + \sigma_T^2(\sigma_y^2 + \mu_y^2)W$ and $R_w = \frac{\mu_y}{\mu_x}$.

- A regression estimator proposed by Gupta et al. (2014) is given by

$$\hat{\mu}_{reg,w} = \bar{z}^* + \hat{\beta}_{zx}(\mu_x - \bar{x}^*), \tag{III.33}$$

where $\hat{\beta}_{zx} = \frac{\sigma_{zx}}{\sigma_x^2} = \rho_{zx}\frac{\sigma_z}{\sigma_x}$.

It can be written as

$$\begin{aligned} \hat{\mu}_{reg,w} &= (1 + e_0^*)\mu_y + \hat{\beta}_{zx}(\mu_x - (1 + e_1^*)\mu_x) \\ &= \mu_y(1 + e_0^*) - \hat{\beta}_{zx}(e_1^*\mu_x). \end{aligned} \tag{III.34}$$

The difference between the regression estimator and the true mean can be written as

$$\hat{\mu}_{reg,w} - \mu_y = e_0^*\mu_y - \hat{\beta}_{zx}(e_1^*\mu_x). \tag{III.35}$$

Taking square and then expected value on both side of (III.35), the MSE of $\hat{\mu}_{reg,w}$, up to second order approximation, is given by

$$MSE^*(\hat{\mu}_{reg,w}) = \theta\sigma_z^2(1 - \rho_{zx}^2) + \theta(\sigma_p^2 + \hat{\beta}_{zx}^2\sigma_v^2). \tag{III.36}$$

The MSEs of the above mean estimators without measurement errors may be obtained by letting $\sigma_p^2 = \sigma_v^2 = 0$ in (III.28), (III.32) and (III.36).

III.4. Generalized Estimator under ORRT Models in the Presence of Measurement Errors

With this background, we use the generalized mean estimator presented in Khalil et al. (2018) [34]. This estimator includes a wide variety of mean estimators as special cases. It is given below:

$$\hat{\mu}_{pw} = (\bar{z}^* + k(\mu_x - \bar{x}^*))(\frac{\bar{D}}{\bar{d}})^v, \tag{III.37}$$

where $\bar{d} = \phi(\alpha\bar{x}+\beta)+(1-\phi)(\alpha\mu_x+\beta)$, $\bar{D} = \alpha\mu_x+\beta$, k and v are suitable constants. ϕ is assumed to be an unknown constant whose value is to be determined from optimally considerations. α ($\alpha \neq 0$) and β are assumed to be some known parameters of the auxiliary variable X, such as coefficient of variation (C_x), kurtosis, and correlation coefficient (ρ_{zx}) etc. Please note here that with different values of α and β, we can obtain various estimators. Also, with v=1 we get various ratio estimators and with v=-1 we get various product estimators.

III.4.1 Bias and MSE of the Generalized Mean Estimator

The generalized mean estimator will be studied under both ORRT model and measurement errors. According to the notations in Chapter III.3, this estimator can be written as

$$\hat{\mu}_{pw} = ((1+e_0^*)\mu_y + k(\mu_x - (1+e_1^*)\mu_x))(\frac{\alpha\mu_x+\beta}{\phi(\alpha(1+e_1^*)\mu_x+\beta)+(1-\phi)(\alpha\mu_x+\beta)})^v. \tag{III.38}$$

Using Taylor's approximation and retaining terms of order up to 2, the difference between the generalized mean estimator and the true mean can be written as

$$
\begin{aligned}
&\hat{\mu}_{pw} - \mu_y \\
&= ((1+e_0^*)\mu_y + k(\mu_x - (1+e_1^*)\mu_x))(\frac{\alpha\mu_x+\beta}{\phi(\alpha(1+e_1^*)\mu_x+\beta)+(1-\phi)(\alpha\mu_x+\beta)})^v - \mu_y \\
&= (\mu_y + e_0^*\mu_y - ke_1^*\mu_x)(\frac{\alpha\mu_x+\beta}{\phi\alpha e_1^*\mu_x+\alpha\mu_x+\beta})^v - \mu_y \\
&\approx \mu_y + (e_0^*-0)\mu_y + (e_1^*-0)[(-k\bar{x}) + \mu_y v(\frac{-\alpha\phi\mu_x}{\alpha\mu_x+\beta})] + \frac{1}{2!}[(e_0^*-0)^2(0) + 2(e_0^*-0) \\
&\quad (e_1^*-0)\mu_y v(\frac{-\alpha\phi\mu_x}{\alpha\mu_x+\beta}) + (e_1^*-0)^2(-k\mu_x v(\frac{-\alpha\phi\mu_x}{\alpha\mu_x+\beta}) - v(-k\mu_x)(\frac{-\alpha\phi\mu_x}{\alpha\mu_x+\beta})+ \\
&\quad v(v-1)\mu_y(\frac{-\alpha\phi\mu_x}{\alpha\mu_x+\beta})(\frac{-\alpha\phi\mu_x}{\alpha\mu_x+\beta}) + 2(\frac{-\alpha\phi\mu_x}{\alpha\mu_x+\beta})(\frac{-\alpha\phi\mu_x}{\alpha\mu_x+\beta})] - \mu_y \\
&= e_0^*\mu_y + e_1^*[-k\mu_x + \mu_y v(\frac{-\alpha\phi\mu_x}{\alpha\mu_x+\beta})] + \frac{1}{2!}[2e_0^*e_1^*\mu_y v(\frac{-\alpha\phi\mu_x}{\alpha\mu_x+\beta}) + e_1^{*2}(2kv\mu_x(\frac{\alpha\phi\mu_x}{\alpha\mu_x+\beta})+ \\
&\quad v(v+1)\mu_y(\frac{-\alpha\phi\mu_x}{\alpha\mu_x+\beta})^2)].
\end{aligned}
\tag{III.39}
$$

Taking expectation on both side of (III.39), the bias of the generalized mean estimator $\hat{\mu}_{pw}$, correct to the second order or approximation, is given by

$$
\begin{aligned}
Bias^*(\hat{\mu}_{pw}) \approx &\frac{\theta}{\mu_y}(\frac{v(v+1)}{2}\phi^2 R_{pw}^2\sigma_x^2 - v\phi R_{pw}\rho_{zx}\sigma_z\sigma_x + v\phi k R_{pw}\sigma_x^2) \\
&+ \frac{\theta}{\mu_y}(\frac{v(v+1)}{2}\phi^2 R_{pw}^2\sigma_v^2 + v\phi k R_{pw}\sigma_v^2),
\end{aligned}
\tag{III.40}
$$

where $R_{pw} = \frac{\alpha\mu_y}{\alpha\mu_x+\beta}$. The bias of $\hat{\mu}_{pw}$ without measurement errors may be obtained by setting $\sigma_v^2 = 0$ in above equation.

To determine the expression for MSE of the generalized mean estimator in (III.37), we take square of (III.39) on both sides and retaining terms of order up to 2 to get

$$(\hat{\mu}_{pw} - \mu_y)^2 = e_0^{*2}\mu_y^2 + k^2\mu_x^2 e_1^{*2} + (\frac{\alpha\phi v}{\alpha\mu_x + \beta}\mu_x\mu_y e_1^*)^2 - 2\rho_{yx}e_1^* k\mu_x\mu_y - \\ 2\rho_{yx}e_1^*\frac{\alpha\phi v}{\alpha\mu_x + \beta}\mu_x\mu_y^2 + 2e_1^{*2}k\frac{\alpha\phi v}{\alpha\mu_x + \beta}\mu_x^2\mu_y. \tag{III.41}$$

Taking the expected value on both side of (III.41), the experssion for MSE of $\hat{\mu}_{pw}$, correct to the first order approximation is given by

$$MSE^*(\hat{\mu}_{pw}) \approx \theta(\sigma_z^2 + v^2\phi^2R_{pw}^2\sigma_x^2 + k^2\sigma_x^2 - 2v\phi R_{pw}\rho_{zx}\sigma_z\sigma_x - 2k\rho_{zx}\sigma_z\sigma_x + 2v\phi kR_{pw}\sigma_x^2) \\ + \theta(\sigma_p^2 + v^2\phi^2R_{pw}^2\sigma_v^2 + k^2\sigma_v^2 + 2v\phi kR_{pw}\sigma_v^2), \tag{III.42}$$

where $\theta = \frac{N-n}{Nn}$ and $R_{pw} = \frac{\alpha\mu_y}{\alpha\mu_x+\beta}$.

The optimum value of ϕ by taking the first derivative which gives the minimum MSE, is given by

$$\phi_{opt} = \frac{\rho_{zx}\sigma_z\sigma_x - k(\sigma_x^2 + \sigma_v^2)}{vR_{pw}(\sigma_x^2 + \sigma_v^2)}. \tag{III.43}$$

By subsitituting (III.43) in (III.42), the minimum value of $MSE^*(\hat{\mu}_{pw})$ is given by

$$MSE_{min}^*(\hat{\mu}_{pw}) \approx \theta(\sigma_z^2 + \sigma_p^2 - \frac{\rho_{zx}^2\sigma_z^2\sigma_x^2}{\sigma_x^2 + \sigma_v^2}). \tag{III.44}$$

The expression for the minimized MSE of proposed estimator without measurement errors may be obtained by putting $\sigma_u^2 = \sigma_v^2 = 0$ in (III.44), which gives

$$MSE_{min}(\hat{\mu}_{pw}) = \theta\sigma_z^2(1 - \rho_{zx}^2). \tag{III.45}$$

III.4.2 Efficiency Comparisons

The $MSE_{min}(\hat{\mu}_{pw})$ in (III.44) is same as that of the approximate MSE of the usual linear regression estimator $MSE(\hat{\mu}_{reg,w})$ (III.36). Comparing the minimum

$MSE(\hat{\mu}_{pw})$ (III.44) with measurement errors to MSEs of existing estimators $MSE(\hat{\mu}_{yw})$ (III.28) and $MSE(\hat{\mu}_{rw})$ (III.32), it is easy to verify that

$$MSE^*_{min}(\hat{\mu}_{pw}) < MSE^*(\hat{\mu}_{yw}) \ \ if \ \ \rho^2_{zx}\frac{\sigma_z^2\sigma_x^2}{\sigma_x^2+\sigma_v^2} > 0 \tag{III.46}$$

and

$$MSE^*_{min}(\hat{\mu}_{pw}) < MSE^*(\hat{\mu}_{rw}) \ \ if \ \ (R_{pw}\sqrt{\sigma_x^2+\sigma_v^2} - \frac{\rho_{zx}\sigma_z\sigma_x}{\sqrt{\sigma_x^2+\sigma_v^2}})^2 > 0. \tag{III.47}$$

These two conditions always hold true.

Comparing $MSE(\hat{\mu}_{yw})$ (III.28) and $MSE(\hat{\mu}_{rw})$ (III.32), it is easy to verify that

$$MSE^*(\hat{\mu}_{rw}) < MSE^*(\hat{\mu}_{yw}) \ \ if \ \ \theta[R_w^2(\sigma_x^2+\sigma_v^2) - 2R_w\rho_{zx}\sigma_x\sigma_z] < 0. \tag{III.48}$$

This condition holds true only when $R_w^2(\sigma_x^2+\sigma_v^2)$ is small or/and $2R_w\rho_{zx}\sigma_x\sigma_z$ is large. The population parameters are fixed for a finite population, and the correlation between X and Z (ρ_{zx}) is positively associated with the correlation between X and Y (ρ_{yx}). In other words, the ratio estimator $\hat{\mu}_{rw}$ is more efficient than the ordinary mean estimator $\hat{\mu}_{yw}$ only when the measurement errors of X (σ_v^2) is small and the correlation between X and Y (ρ_{yx}) is high. This conditional superiority of the ratio estimator is reasonable because measurement errors on both X and Y will bring more burden on efficiency than the measurement errors on Y alone. However, the generalized mean estimator has no such restrictions. It is always more efficient than the ordinary mean estimator no matter how large the measurement errors are sine the generalized estimator uses variety of other information also.

III.5. Simulation Study

In this section, we examine the performance of the generalized mean estimator with the ordinary mean estimator and the ratio estimator, by way of a simulation study. In the generalized mean estimator, we choose v and k to be 1, and ϕ to be its optimum value. As for α and β, we have used various parameters associated with the auxiliary variable such as the coefficient of variation (C_x) and kurtosis, but these choices do not impact the results in any meaningful way. As we can see in (III.44) above, minimized MSE is independent of α and β, and empirical MSEs also are almost the same for all choices of α and β. We will show different cases where α and β take different values.

We consider a finite population of size 5000 generated from bivariate normal distribution with means and covariances of (Y, X) as given below.

$$\textbf{Population} \quad \mu = \begin{bmatrix} 10 \\ 6 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 16 & 9.0510 \\ 9.0510 & 8 \end{bmatrix}, \quad \rho_{yx} = 0.8$$

To explain the simulation process further, we started with a sample of size 5000 from a normal population with parameters:

$$\mu_x = 6,\ \sigma_x^2 = 8,\ \mu_y = 10,\ \sigma_y^2 = 16,\ \rho_{yx} = 0.8 \tag{A}$$

However, the real parameters of the set of 5000 data points we generated using R are very close to the parameter values in (A) but not exactly same. For the simulation study, we used parameter values in (B) and not those in (A).

$$\mu_x = 6.0228,\ \sigma_x^2 = 8.1830,\ \mu_y = 9.9864,\ \sigma_y^2 = 16.1215,\ \rho_{yx} = 0.8024 \tag{B}$$

The scrambling variable S is taken to be a normal variate with mean equal to zero and vary variances ($0.2^*\sigma_x^2$, $0.5^*\sigma_x^2$ and $1^*\sigma_x^2$). And T is also taken to be a normal variate but with mean equal to one and varying variances (0, 0.5, 1).

The observed values of Z and X are given by: $z = Z + p$ and $x = X + v$, where p and v are represent measurement errors. The measurement errors are taken to be a normal distributions with mean equal to zero and varying variances (0, 5, 10). The observed response z is given by z=TY+S+P with probability W, and by z = Y+P with probability 1-W.

We consider samples of size n = 500 using SRSWOR (simple random sampling without replacement). Coding for the simulations was done in R and results are averaged over 5,000 iterations. The empirical MSE of the estimator $\hat{\mu}_w$ is computed by

$$MSE^*(\hat{\mu}_w) = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\mu}_w - \mu_y)^2, \tag{III.49}$$

where $\hat{\mu}_w = \hat{\mu}_{yw}, \hat{\mu}_{rw}, \hat{\mu}_{pw}$. Here, μ_y is the population mean of the sensitive study variable. The percent relative efficiencies (PREs) of the estimators ($\hat{\mu}_w$) with respect to mean estimator ($\hat{\mu}_{yw}$) is defined as

$$PRE = \frac{MSE^*(\hat{\mu}_{yw})}{MSE^*(\hat{\mu}_w)} * 100. \tag{III.50}$$

We will also use the unified measure δ of the efficiency and the privacy as defined in Gupta et al. (2018)[24]. It is given by

$$\delta = \frac{MSE^*(\hat{\mu}_w)}{\Delta_{DP}}. \tag{III.51}$$

In (III.51), MSE is used in place of Var(.) to account for biased estimators.

Table III.1. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators when $\sigma_v^2 = \sigma_p^2 = 1$ and $\sigma_s^2 = 0.2^*\sigma_x^2$.

Est.	W	σ_T^2	MSE		PRE		δ
			Without ME	With ME	Without ME	With ME	
$\hat{\mu}_{yw}$	0.5	0	**0.0305**	**0.0323**	**100.0000**	**100.0000**	**0.0198**
			0.0311	0.0335	100.0000	100.0000	0.0205
		0.5	**0.0820**	**0.0838**	**100.0000**	**100.0000**	**0.0014**
			0.0822	0.0884	100.0000	100.0000	0.0015
		1	**0.1335**	**0.1352**	**100.0000**	**100.0000**	**0.0012**
			0.1330	0.1340	100.0000	100.0000	0.0012
	0.8	0	**0.0313**	**0.0331**	**100.0000**	**100.0000**	**0.0202**
			0.0309	0.0334	100.0000	100.0000	0.0204
		0.5	**0.1137**	**0.1155**	**100.0000**	**100.0000**	**0.0020**
			0.1120	0.1130	100.0000	100.0000	0.0019
		1	**0.1961**	**0.1979**	**100.0000**	**100.0000**	**0.0017**
			0.1930	0.1933	100.0000	100.0000	0.0017
	1	0	**0.0319**	**0.0337**	**100.0000**	**100.0000**	**0.0206**
			0.0312	0.0339	100.0000	100.0000	0.0207
		0.5	**0.1349**	**0.1367**	**100.0000**	**100.0000**	**0.0023**
			0.1334	0.1351	100.0000	100.0000	0.0023
		1	**0.2379**	**0.2397**	**100.0000**	**100.0000**	**0.0021**
			0.2350	0.2368	100.0000	100.0000	0.0020
$\hat{\mu}_{rw}$	0.5	0	**0.0160**	**0.0227**	**190.6250**	**142.2907**	**0.0139**
			0.0165	0.0231	188.4848	145.0216	0.0141
		0.5	**0.0675**	**0.0742**	**121.4815**	**112.9380**	**0.0013**
			0.0678	0.0794	121.2389	111.3350	0.0013
		1	**0.1189**	**0.1257**	**112.2792**	**107.5577**	**0.0011**
			0.1286	0.1249	103.4215	107.2858	0.0011
	0.8	0	**0.0168**	**0.0236**	**186.3095**	**140.2542**	**0.0144**
			0.0174	0.0241	177.5862	138.5892	0.0147
		0.5	**0.0992**	**0.1060**	**114.6169**	**108.9623**	**0.0018**
			0.1006	0.1055	111.3320	107.1090	0.0018

		1	**0.1816**	**0.1883**	**107.9846**	**105.0982**	**0.0016**
			0.1826	0.1863	105.6955	103.7574	0.0016
	1	0	**0.0174**	**0.0242**	**183.3333**	**139.2562**	**0.0148**
			0.0180	0.0246	173.3333	137.8049	0.0150
		0.5	**0.1204**	**0.1271**	**112.0432**	**107.5531**	**0.0022**
			0.1219	0.1281	109.4340	105.4645	0.0022
		1	**0.2234**	**0.2301**	**106.4906**	**104.1721**	**0.0020**
			0.2249	0.2307	104.4909	102.6441	0.0020
$\hat{\mu}_{pw}$	0.5	0	**0.0118**	**0.0156**	**258.4746**	**207.0513**	**0.0095**
			0.0119	0.0160	261.3445	209.3750	0.0098
		0.5	**0.0633**	**0.0671**	**129.5419**	**124.8882**	**0.0011**
			0.0682	0.0678	120.5279	130.3835	0.0012
		1	**0.1148**	**0.1186**	**116.2892**	**113.9966**	**0.0010**
			0.1140	0.1174	116.6667	114.1397	0.0010
	0.8	0	**0.0127**	**0.0165**	**246.4567**	**200.6061**	**0.0101**
			0.0129	0.0170	239.5349	196.4706	0.0104
		0.5	**0.0951**	**0.0989**	**119.5584**	**116.7846**	**0.0017**
			0.0954	0.0977	117.4004	115.6602	0.0017
		1	**0.1774**	**0.1813**	**110.5411**	**109.1561**	**0.0016**
			0.1771	0.1784	108.9780	108.3520	0.0015
	1	0	**0.0132**	**0.0171**	**241.6667**	**197.0760**	**0.0105**
			0.0135	0.0173	231.1111	195.9538	0.0106
		0.5	**0.1163**	**0.1201**	**115.9931**	**113.8218**	**0.0020**
			0.1169	0.1202	114.1146	112.3960	0.0020
		1	**0.2192**	**0.2230**	**108.5310**	**107.4888**	**0.0019**
			0.2195	0.2225	107.0615	106.4270	0.0019

Table III.2. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators when $\sigma_v^2 = \sigma_p^2 = 1$ and $\sigma_s^2 = 0.5*\sigma_x^2$.

Est.	W	σ_T^2	MSE Without ME	MSE With ME	PRE Without ME	PRE With ME	δ
$\hat{\mu}_{yw}$	0.5	0	**0.0327**	**0.0344**	**100.0000**	**100.0000**	**0.0084**
			0.0335	0.0360	100.0000	100.0000	0.0088
		0.5	**0.0842**	**0.0860**	**100.0000**	**100.0000**	**0.0014**
			0.0895	0.0908	100.0000	100.0000	0.0015
		1	**0.1357**	**0.1375**	**100.0000**	**100.0000**	**0.0012**
			0.1354	0.1464	100.0000	100.0000	0.0012
	0.8	0	**0.0348**	**0.0366**	**100.0000**	**100.0000**	**0.0090**
			0.0344	0.0371	100.0000	100.0000	0.0091
		0.5	**0.1173**	**0.1191**	**100.0000**	**100.0000**	**0.0019**
			0.1157	0.1167	100.0000	100.0000	0.0019
		1	**0.1997**	**0.2014**	**100.0000**	**100.0000**	**0.0017**
			0.1967	0.1971	100.0000	100.0000	0.0017
	1	0	**0.0363**	**0.0381**	**100.0000**	**100.0000**	**0.0093**
			0.0355	0.0383	100.0000	100.0000	0.0094
		0.5	**0.1393**	**0.1411**	**100.0000**	**100.0000**	**0.0023**
			0.1387	0.1403	100.0000	100.0000	0.0023
		1	**0.2423**	**0.2441**	**100.0000**	**100.0000**	**0.0021**
			0.2405	0.2421	100.0000	100.0000	0.0020
$\hat{\mu}_{rw}$	0.5	0	**0.0181**	**0.0249**	**180.6630**	**138.1526**	**0.0061**
			0.0189	0.0256	177.2487	140.6250	0.0063
		0.5	**0.0697**	**0.0764**	**120.8034**	**112.5654**	**0.0012**
			0.0702	0.0818	127.4929	111.0024	0.0013
		1	**0.1212**	**0.1279**	**111.9637**	**107.5059**	**0.0011**
			0.1211	0.1274	111.8084	114.9137	0.0011
	0.8	0	**0.0203**	**0.0271**	**171.4286**	**135.0554**	**0.0066**
			0.0212	0.0280	162.2642	132.5000	0.0069
		0.5	**0.1028**	**0.1095**	**114.1051**	**108.7671**	**0.0018**
			0.1044	0.1092	110.8238	106.8681	0.0018

		1	**0.1851**	**0.1919**	**107.8876**	**104.9505**	**0.0016**
			0.1864	0.1900	105.5258	103.7368	0.0016
	1	0	**0.0218**	**0.0285**	**166.5138**	**133.6842**	**0.0070**
			0.0227	0.0293	156.3877	130.7167	0.0072
		0.5	**0.1248**	**0.1316**	**111.6186**	**107.2188**	**0.0021**
			0.1268	0.1328	109.3849	105.6476	0.0022
		1	**0.2278**	**0.2345**	**106.3652**	**104.0938**	**0.0020**
			0.2301	0.2356	104.5198	102.7589	0.0020
$\hat{\mu}_{pw}$	0.5	0	**0.0140**	**0.0178**	**233.5714**	**193.2584**	**0.0044**
			0.0143	0.0184	234.2657	195.6522	0.0045
		0.5	**0.0655**	**0.0693**	**128.5496**	**124.0981**	**0.0011**
			0.0706	0.0721	126.7705	125.9362	0.0012
		1	**0.1170**	**0.1208**	**115.9829**	**113.8245**	**0.0010**
			0.1164	0.1298	116.3230	112.7889	0.0011
	0.8	0	**0.0162**	**0.0200**	**214.8148**	**183.0000**	**0.0049**
			0.0167	0.0208	205.9880	178.3654	0.0051
		0.5	**0.0986**	**0.1024**	**118.9655**	**116.3086**	**0.0017**
			0.0992	0.1014	116.6331	115.0888	0.0017
		1	**0.1810**	**0.1848**	**110.3315**	**108.9827**	**0.0016**
			0.1808	0.1821	108.7942	108.2372	0.0015
	1	0	**0.0176**	**0.0214**	**206.2500**	**178.0374**	**0.0052**
			0.0181	0.0219	196.1326	174.8858	0.0054
		0.5	**0.1207**	**0.1245**	**115.4101**	**113.3333**	**0.0020**
			0.1219	0.1250	113.7818	112.2400	0.0020
		1	**0.2236**	**0.2275**	**108.3631**	**107.2967**	**0.0019**
			0.2247	0.2276	107.0316	106.3708	0.0019

Table III.3. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators when $\sigma_v^2 = \sigma_p^2 = 1$ and $\sigma_s^2 = 1{*}\sigma_x^2$.

Est.	W	σ_T^2	MSE Without ME	MSE With ME	PRE Without ME	PRE With ME	δ
$\hat{\mu}_{yw}$	0.5	0	**0.0363**	**0.0381**	**100.0000**	**100.0000**	**0.0047**
			0.0376	0.0400	100.0000	100.0000	0.0049
		0.5	**0.0879**	**0.0896**	**100.0000**	**100.0000**	**0.0014**
			0.0875	0.0907	100.0000	100.0000	0.0014
		1	**0.1393**	**0.1411**	**100.0000**	**100.0000**	**0.0012**
			0.1393	0.1404	100.0000	100.0000	0.0011
	0.8	0	**0.0407**	**0.0424**	**100.0000**	**100.0000**	**0.0052**
			0.0404	0.0432	100.0000	100.0000	0.0053
		0.5	**0.1232**	**0.1250**	**100.0000**	**100.0000**	**0.0019**
			0.1219	0.1229	100.0000	100.0000	0.0019
		1	**0.2055**	**0.2073**	**100.0000**	**100.0000**	**0.0017**
			0.2028	0.2032	100.0000	100.0000	0.0017
	1	0	**0.0436**	**0.0454**	**100.0000**	**100.0000**	**0.0056**
			0.0428	0.0458	100.0000	100.0000	0.0057
		0.5	**0.1467**	**0.1485**	**100.0000**	**100.0000**	**0.0023**
			0.1470	0.1485	100.0000	100.0000	0.0023
		1	**0.2497**	**0.2515**	**100.0000**	**100.0000**	**0.0021**
			0.2491	0.2506	100.0000	100.0000	0.0020
$\hat{\mu}_{rw}$	0.5	0	**0.0218**	**0.0285**	**166.5138**	**133.6842**	**0.0035**
			0.0219	0.0296	171.6895	135.1351	0.0037
		0.5	**0.0733**	**0.0801**	**119.9181**	**111.8602**	**0.0012**
			0.0793	0.0858	110.3405	105.7110	0.0013
		1	**0.1248**	**0.1316**	**111.6186**	**107.2188**	**0.0011**
			0.1251	0.1314	111.3509	106.8493	0.0011
	0.8	0	**0.0261**	**0.0329**	**155.9387**	**128.8754**	**0.0041**
			0.0275	0.0343	146.9091	125.9475	0.0042
		0.5	**0.1087**	**0.1154**	**113.3395**	**108.3189**	**0.0018**
			0.1077	0.1152	113.1848	106.6840	0.0018

		1	**0.1910**	**0.1977**	**107.5916**	**104.8558**	**0.0016**
			0.1927	0.1961	105.2413	103.6206	0.0016
	1	0	**0.0291**	**0.0358**	**149.8282**	**126.8156**	**0.0044**
			0.0306	0.0371	139.8693	123.4501	0.0046
		0.5	**0.1322**	**0.1389**	**110.9682**	**106.9114**	**0.0021**
			0.1348	0.1406	109.0504	105.6188	0.0022
		1	**0.2382**	**0.2419**	**104.8279**	**103.9686**	**0.0020**
			0.2352	0.2435	105.9099	102.9158	0.0020
$\hat{\mu}_{pw}$	0.5	0	**0.0176**	**0.0214**	**206.2500**	**178.0374**	**0.0026**
			0.0183	0.0224	205.4645	178.5714	0.0028
		0.5	**0.0692**	**0.0730**	**127.0231**	**122.7397**	**0.0011**
			0.0686	0.0731	127.5510	124.0766	0.0011
		1	**0.1207**	**0.1245**	**115.4101**	**113.3333**	**0.0010**
			0.1204	0.1238	115.6977	113.4087	0.0010
	0.8	0	**0.0220**	**0.0258**	**185.0000**	**164.3411**	**0.0032**
			0.0228	0.0270	177.1930	160.0000	0.0033
		0.5	**0.1045**	**0.1083**	**117.8947**	**115.4201**	**0.0017**
			0.1054	0.1075	115.6546	114.3256	0.0016
		1	**0.1869**	**0.1907**	**109.9518**	**108.7048**	**0.0016**
			0.1870	0.1882	108.4492	107.9702	0.0015
	1	0	**0.0249**	**0.0287**	**175.1004**	**158.1882**	**0.0036**
			0.0258	0.0295	165.8915	155.2542	0.0037
		0.5	**0.1280**	**0.1318**	**114.6094**	**112.6707**	**0.0020**
			0.1300	0.1330	113.0769	111.6541	0.0020
		1	**0.2310**	**0.2348**	**108.0952**	**107.1124**	**0.0019**
			0.2330	0.2358	106.9099	106.2765	0.0019

Tables III.1, III.2, and III.3 present the theoretical and empirical MSEs and PREs of the ORRT mean estimators when both the variances of measurement errors on X and Z are set equal to 1 and the variance of S is set equal to 0.2*σ_x^2, 0.5*σ_x^2 and 1*σ_x^2, respectively. Comparing these three tables, the mean estimation is less efficient as the variance of S increases. For instance, when σ_T^2=1, the sensitively level W is is equal to 0.8, and the measurement errors are present, the MSEs of the generalized mean estimator are 0.1813, 0.1848, 0.1907 respectively corresponding to the Var(S) is equal to 0.2*σ_x^2, 0.5*σ_x^2, and $1 * \sigma_x^2$. These results consistent with the theoretical results. Larger variance of S introduces more penalty for using RRT models.

From all three tables, one can observe that the MSE of the mean estimators increases as W increases, both when measurement errors are present, and absent. For example in the Table III.1, the MSE of the generalized mean estimator increased from 0.0671 to 0.1201 as W increased from 0.5 to 1 when $\sigma_T^2 = 0.5$ and the measurement errors are present. It indicates the ORRT model gains some efficiency when some the respondents feel the survey question is not sensitive. Also, as the variance of T increases, the MSE increases while δ decreases with a reasonably small value of σ_T^2. Again, we can select some values from Table III.1 as an example. When the sensitively level W is 0.5 and the measurement errors are present, the MSE of the ordinary mean estimator increased from 0.0323 to 0.1352 as the variance of T increased from 0 to 1, while the δ value decreased from 0.0198 to 0.0012. In other words, mean estimators under the simple additive model (Z=Y+S) are more efficient as compared to the general linear combination model (Z=TY+S). However, the general linear combination model is better if both efficiency and privacy are considered simultaneously.

Table III.4. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators under the Conditions of $\sigma_v^2 = \sigma_p^2 = 1$, 5, 10 when W = 0.8, $\sigma_T^2 = 0.5$ and $\sigma_s^2 = 0.5{*}\sigma_x^2$.

	MSE			PRE		
Est.	1	5	10	1	5	10
$\hat{\mu}_{yw}$	**0.1191**	**0.1262**	**0.1351**	**100.0000**	**100.0000**	**100.0000**
	0.1167	0.1241	0.1335	100.0000	100.0000	100.0000
$\hat{\mu}_{rw}$	**0.1095**	**0.1364**	**0.1699**	**108.7671**	**92.5220**	**79.5174**
	0.1092	0.1359	0.1703	106.8681	91.3171	78.3911
$\hat{\mu}_{pw}$	**0.1024**	**0.1146**	**0.1267**	**116.3086**	**110.1222**	**106.6298**
	0.1014	0.1133	0.1260	115.0888	109.5322	105.9524

Table III.4 presents the theoretical and empirical MSEs and PREs of the ORRT mean estimators under different variances of measurement errors on X and Z when the sensitivity level W is equal to 0.8, variance of T is equal to 0.5 and variance of S is equal to $0.5{*}\sigma_x^2$. As the variance of measurement errors increase, the MSE of each mean estimator increases, which means larger measurement errors have larger negative impact on mean estimation.

Also, from Tables III.1, III.2, III.3 and III.4, it is more clear that the generalized mean estimator $\hat{\mu}_{pw}$ is more efficient than the other two mean estimators no matter how large the measurement errors are. However, as the measurement errors increase, the ratio estimator $\hat{\mu}_{rw}$ become less efficient than the ordinary mean estimator $\hat{\mu}_{yw}$ because the ordinary mean estimator is not impacted by the measurement error in X. This was not so for the generalized mean estimator because the use of the regression term was able to overcome the measurement error burden due to X. Therefore, the generalized mean estimator may be preferred in mean estimation since it is more efficient without restrictions.

Table III.5. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Generalized Mean Estimator under Different α and β Values when $\sigma_v^2 = \sigma_p^2 = 1$, W = 0.8 and $\sigma_s^2 = 0.5^*\sigma_x^2$.

α	β	σ_T^2	MSE Without ME	MSE With ME	PRE Without ME	PRE With ME	δ
1	0	0	**0.0162**	**0.0200**	**214.8148**	**118.9189**	**0.0109**
			0.0167	0.0208	205.9880	118.6975	0.0116
		0.5	**0.0986**	**0.1024**	**118.9655**	**106.6298**	**0.0021**
			0.0992	0.1014	116.6331	105.9524	0.0021
		1	**0.1810**	**0.1848**	**110.3315**	**104.0172**	**0.0018**
			0.1808	0.1821	108.7942	103.5351	0.0017
1	C_x	0	**0.0162**	**0.0200**	**214.8148**	**118.9189**	**0.0109**
			0.0167	0.0208	205.9880	118.6975	0.0116
		0.5	**0.0986**	**0.1024**	**118.9655**	**106.6298**	**0.0021**
			0.0992	0.1014	116.6331	105.9524	0.0021
		1	**0.1810**	**0.1848**	**110.3315**	**104.0172**	**0.0018**
			0.1808	0.1821	108.7942	103.5351	0.0017
0.5	0	0	**0.0162**	**0.0200**	**214.8148**	**118.9189**	**0.0109**
			0.0167	0.0208	205.9880	118.6975	0.0116
		0.5	**0.0986**	**0.1024**	**118.9655**	**106.6298**	**0.0021**
			0.0992	0.1014	116.6331	105.9524	0.0021
		1	**0.1810**	**0.1848**	**110.3315**	**104.0172**	**0.0018**
			0.1808	0.1821	108.7942	103.5351	0.0017
0.5	C_x	0	**0.0162**	**0.0200**	**214.8148**	**118.9189**	**0.0109**
			0.0167	0.0208	205.9880	118.6975	0.0116
		0.5	**0.0986**	**0.1024**	**118.9655**	**106.6298**	**0.0021**
			0.0992	0.1014	116.6331	105.9524	0.0021
		1	**0.1810**	**0.1848**	**110.3315**	**104.0172**	**0.0018**
			0.1808	0.1821	108.7942	103.5351	0.0017

Table III.5 presents the theoretical and empirical MSEs and PREs of the ORRT estimators under different α and β values when the sensitivity level W is equal to 0.8

and variance of T is equal to 0.5. It is clear that different values of α and β have no impact on the efficiency.

III.6. Concluding Chapter Remarks

The main contribution in this chapter is the mean estimation of a sensitive variable, in the presence of measurement errors, using ORRT models. While such mean estimation has been attempted before by Khalil et al. (2018) using non-optional RRT models. It has not been done using the more efficient ORRT models. The resultant gain in efficiency using ORRT models is obvious from both the theoretical and the empirical results. The simple additive RRT model is more efficient in terms of PRE. But the general RRT model is better, if we examine the performance of various estimators with respect to the unified measure of efficiency and privacy. It is also clear from the theoretical conditions (III.46) and (III.47) and the simulation results that the generalized mean estimator is more efficient than the ordinary mean estimator and the ratio estimator.

Non-response is another common non-sampling error we have seen in sampling. Will the generalized estimator in Chapter III still be more efficient than the other existing estimators in the presence of non-response? The mean estimation of a sensitive variable under both measurement errors and non-response will be discussed in the next chapter.

CHAPTER IV

MEAN ESTIMATION IN THE SIMULTANEOUS PRESENCE OF MEASUREMENT ERRORS AND NON-RESPONSE USING ORRT MODELS

We have briefly introduced utilizing non-optional RRT in Hansen and Hurwitz (1946) two-phase sampling in Section II.3. But we aim to work with the more efficient ORRT models. In Section IV.1, a modified version of Hansen and Hurwitz (1946) two-phase sampling using ORRT models will be introduced[91]; some existing mean estimators under the modified two-phase sampling in the presence of measurement errors will be discussed in Section IV.2; Section IV.3 will talk about the generalized mean estimator; Section IV.4 will present the simulation results; Section IV.5 will provide concluding remarks for this Chapter.

IV.1. Modified Hansen and Hurwitz (HH) Two-phase Sampling Technique

As mentioned in Section I.2, Hansen and Hurwitz two-phase sampling uses mail or phone survey at the first attempt and then uses face-to-face interview at the second phase to obtain more information. However, it may cause non-response bias if the variable of interest is sensitive. The respondent may provide untruthful response in the face-to-face interview. In order to encourage the respondents to answer a sensitive survey question truthfully, we give the respondents the opportunity to scramble the response using ORRT in the second phase of HH procedure when there is a face-to-face interview. In this case, we are modifying the HH procedure assuming that in the first phase, respondent group gives direct answer to both X and Y; and then in the second phase, ORRT model is used to get response from the group of non-respondents.

Using the standard terminology as used before, Let $\mu_y = \frac{\sum_{i=1}^{N} y_i}{N}$ and $\sigma_y^2 = \frac{\sum_{i=1}^{N}(y_i-\mu_x)^2}{N-1}$ be the population mean and variance of the study variable Y. Let $\mu_{y_{(1)}} = \frac{\sum_{i=1}^{N_1} y_i}{N_1}$ and $\sigma_{y_{(1)}}^2 = \frac{\sum_{i=1}^{N_1}(y_i-\mu_{y_1})^2}{N_1-1}$ be the population mean and variance of respondent group of size N_1, $\mu_{y_{(2)}} = \frac{\sum_{i=1}^{N_2} y_i}{N_2}$ and $\sigma_{y_{(2)}}^2 = \frac{\sum_{i=1}^{N_2}(y_i-\mu_{y_2})^2}{N_2-1}$ be the population mean and variance of non-respondent group of size N_2. Let $\mu_x = \frac{\sum_{i=1}^{N} x_i}{N}$ and $\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i-\mu_x)^2}{N-1}$ be the population mean and variance of the auxiliary variable X. Let $\mu_{x_{(1)}} = \frac{\sum_{i=1}^{N_1} x_i}{N_1}$ and $\sigma_{x_{(1)}}^2 = \frac{\sum_{i=1}^{N_1}(x_i-\mu_{x_1})^2}{N_1-1}$ be the population mean and variance of respondent group of size N_1, $\mu_{x_{(2)}} = \frac{\sum_{i=1}^{N_2} x_i}{N_2}$ and $\sigma_{x_{(2)}}^2 = \frac{\sum_{i=1}^{N_2}(x_i-\mu_{x_2})^2}{N_2-1}$ be the population mean and variance of non-respondent group of size N_2. We assume that only n_1 units provide response on the first call and remaining $n_2 = n - n_1$ units do not respond. Then a subsample of size $n_s = \frac{n_2}{f}$ (f>0) is taken. Let $\rho_{yx} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ be the correlation coefficient between X and Y. Similarly let $\rho_{xy_{(1)}} = \frac{\sigma_{xy_{(1)}}}{\sigma_x \sigma_y}$ be the correlation coefficient for the respondent group, and $\rho_{xy_{(2)}} = \frac{\sigma_{xy_{(2)}}}{\sigma_x \sigma_y}$ be the correlation coefficient for the non-respondents group.

In Section III.2, we have proved that the general linear combination RRT model is better if both efficiency and privacy are considered together. Therefore, when we apply ORRT in the second phase, the scrambled response is given by

$$Z = \begin{cases} Y & \text{with probability 1-W} \\ TY + S & \text{with probability W,} \end{cases} \tag{IV.1}$$

where it is assumed that μ_T =E(T)=1 and μ_s =E(S)=0.

We can write randomized linear model as Z=(YT+S)J+Y(1-J), where J $\sim$ Bernoulli(W). Therefore, E(J)=W, Var(J)=W(1-W) and E(J^2)=Var(J)+ E^2(J)=W.

The expectation and variance under randomization mechanism are given by

$$\begin{aligned} E_R(Z) &= E_R(TYJ + SJ + Y - YJ) \\ &= YE_R(TJ) + E_R(SJ) + Y - YE_R(J) \\ &= Y\mu_T W + \mu_s W + Y - YW \\ &= (\mu_T W + 1 - W)Y + \mu_s W \end{aligned} \tag{IV.2}$$

and

$$\begin{aligned} V_R(Z) &= V_R(TYJ + SJ + Y - YJ) \\ &= V_R(TYJ) + V_R(SJ) + V_R(YJ) + 2Cov(TYJ, SJ) - 2Cov(TYJ, YJ) \\ &\quad - 2Cov(SJ, YJ) \\ &= Y^2[(\sigma_T^2 + \mu_T^2)W - \mu_T^2 W^2] + [(\sigma_s^2 + \mu_s^2)W - \mu_s^2 W^2] + Y^2[W(1-W)+ \\ &\quad 2Y\mu_T\mu_s W(1-W) - 2Y^2[\mu_T W(1-W)] - 2Y[\mu_s W(1-W)] \\ &= (Y^2\sigma_T^2 + \sigma_s^2)W. \end{aligned} \tag{IV.3}$$

Let $\hat{y}_i$ be a transformation of the randomized response on the i^{th} unit whose expectation under the randomization mechanism is the true response y_i. It is given by

$$\hat{y}_i = \frac{z_i - \mu_s W}{\mu_T W + 1 - W} \tag{IV.4}$$

with

$$E_R(\hat{y}_i) = y_i \tag{IV.5}$$

(from IV.2) and

$$V_R(\hat{y}_i) = \frac{V_R(z_i)}{(\mu_T W + 1 - W)^2} \\ = \frac{[y_i^2\sigma_T^2 + \sigma_s^2]W}{(\mu_T W + 1 - W)^2} = \tau_i \tag{IV.6}$$

(from IV.3).

With ORRT model added, a modified version of HH estimator is given by

$$\hat{\bar{y}} = w_1\bar{y}_1 + w_2\hat{\bar{y}}_2, \tag{IV.7}$$

where $\hat{\bar{y}}_2 = \sum_{i=1}^{n_s}(\frac{\hat{y}_i}{n_s})$.

Let E_i and V_i be the expectation and variance in the i^{th} phase (i=1,2) under the two-phase sampling. It is easy to verify that

$$\begin{aligned} E(\hat{\bar{y}}) &= E_1E_2[w_1\bar{y}_1 + w_2\hat{\bar{y}}_2] \\ &= E_1[w_1\bar{y}_1 + w_2E_R(\hat{\bar{y}}_2)] \\ &= E_1[w_1\bar{y}_1 + w_2\bar{y}_2)] \\ &= W_1\mu_{y_{(1)}} + W_2\mu_{y_{(2)}} \\ &= \mu_y \end{aligned} \tag{IV.8}$$

since $E_R(\hat{\bar{y}}_2) = \frac{1}{n_s}\sum_{i=1}^{n_s} E_R(\hat{y}_i) = \bar{y}_2$.

The variance of $\hat{\hat{y}}$ can be written as

$$\begin{aligned}
Var(\hat{\hat{y}}) &= E_1[V_2(\hat{\hat{y}})] + V_1[E_2(\hat{\hat{y}})] \\
&= E_1[V_2(w_1\bar{y}_1 + w_2\hat{\bar{y}}_2)] + V_1[E_2(w_1\bar{y}_1 + w_2\hat{\bar{y}}_2)] \\
&= E_1[0 + V_2(w_2\hat{\bar{y}}_2)] + V_1[w_1\bar{y}_1 + w_2\bar{y}_2] \\
&= E_1[V_2(w_2\hat{\bar{y}}_2)] + V_1(\bar{y}) \\
&= E_1[\frac{w_2^2}{n_s}\frac{\sum_{i=1}^{N_2}\frac{(y_i^2\sigma_T^2+\sigma_s^2)W}{(\mu_T W+1-W)^2}}{N_2}] + V(\bar{y}) \\
&= Var(\bar{y}) + \frac{W_2 f}{n}\frac{\sum_{i=1}^{N_2}\tau_i}{N_2}.
\end{aligned} \tag{IV.9}$$

Note $E(y_i^2) = \sigma_y^2 + \mu_y^2$, and

$$E(\frac{w_2^2}{n_s}) = E(\frac{n_2^2}{n^2}\frac{f}{n_2}) = E(\frac{n_2 f}{n^2}) = \frac{f}{n^2}E(n_2) = \frac{f}{n^2}(nW_2) = \frac{W_2 f}{n}, \tag{IV.10}$$

if we assume $\frac{n}{N} \approx \frac{n_2}{N_2}$.

Since $\bar{y}$ is the original HH mean estimator, the variance of $\hat{\hat{y}}$ is given by

$$Var(\hat{\hat{y}}) = \theta\sigma_y^2 + \lambda\sigma_{y_{(2)}}^2 + \frac{W_2 f}{n}[\frac{[(\sigma_{y_{(2)}}^2 + \mu_{y_{(2)}}^2)\sigma_T^2 + \sigma_s^2]W}{(\mu_T W + 1 - W)^2}], \tag{IV.11}$$

where $\theta = \frac{(N-n)}{Nn}$ and $\lambda = \frac{(f-1)W_2}{n}$. It is easy to notice that $\frac{W_2 f}{n}[\frac{[(\sigma_{y_{(2)}}^2+\mu_{y_{(2)}}^2)\sigma_T^2+\sigma_s^2]W}{(\mu_T W+1-W)^2}]$ is the penalty for using ORRT model.

IV.2. Some Existing Mean Estimators under Measurement Errors and Non-response

Let the measurement error of the auxiliary variable (X) in the population be given by $V_i = x_i - X_i$. Let the respective measurement errors associated with the study variable (Y) in the population and the scrambled variable (Z) in the face-to-face phase be given by $U_i = y_i - Y_i$ and $P_i = z_i - Z_i$. These measurement errors are assumed to be random and uncorrelated with mean zero and variances σ_v^2, σ_u^2, and σ_p^2, respectively.

Assume population mean μ_x of the auxiliary variable X is known, and non-response happens on both X and Y. Some notations are given below

$$\Omega_y = \sum_{i=1}^{n}(y_i - \mu_y), \tag{IV.12}$$

$$\Omega_x = \sum_{i=1}^{n}(x_i - \mu_x), \tag{IV.13}$$

$$\Omega_u = \sum_{i=1}^{n_1} U_i + \sum_{i=1}^{n_2} P_i, \tag{IV.14}$$

and

$$\Omega_v = \sum_{i=1}^{n} V_i, \tag{IV.15}$$

where U_i, P_i, V_i are measurement errors on Y, Z and X, respectively. Let $e_0^* = \frac{1}{n\mu_y}(\Omega_y + \Omega_u)$ and $e_1^* = \frac{1}{n\mu_x}(\Omega_x + \Omega_v)$. In other words, $\hat{\bar{y}}^* = (1 + e_0^*)\mu_y$ and $\bar{x}^* = (1 + e_1^*)\mu_x$, where $\hat{\bar{y}}^* = w_1\bar{y}_1^* + w_2\hat{\bar{y}}_2^*$ and $\bar{x}^* = w_1\bar{x}_1^* + w_2\bar{x}_2^*$ in the presence of measurement errors.

Under the assumption of bivariate normality (Sukhatme et al.1970)[78]:

$$E(e_0^*) = E(e_1^*) = 0; \tag{IV.16}$$

$$E(e_0^{*2}) = \frac{1}{\mu_y^2}\{\theta(\sigma_y^2 + \sigma_u^2) + \lambda(\sigma_{y_{(2)}}^2 + \sigma_p^2) + \frac{W_2 f}{n}[\frac{[(\sigma_{y_{(2)}}^2 + \mu_{y_{(2)}}^2)\sigma_T^2 + \sigma_s^2]W}{(\mu_T W + 1 - W)^2}]\}; \tag{IV.17}$$

$$E(e_1^{*2}) = \frac{1}{\mu_x^2}[\theta(\sigma_x^2 + \sigma_v^2) + \lambda(\sigma_{x(2)}^2 + \sigma_v^2)]; \tag{IV.18}$$

and

$$E(e_0^* e_1^*) = \theta \rho_{yx} \frac{\sigma_y}{\mu_y} \frac{\sigma_x}{\mu_x} + \lambda \rho_{zx(2)} \frac{\sigma_z}{\mu_y} \frac{\sigma_{x(2)}}{\mu_x}, \tag{IV.19}$$

where $\theta = \frac{(N-n)}{Nn}$, $\lambda = \frac{(f-1)W_2}{n}$, $\sigma_z^2 = \sigma_y^2 + \sigma_s^2 W + \sigma_T^2(\sigma_y^2 + \mu_y^2)W$, and

$$\rho_{zx_{(2)}} = \frac{\rho_{yx(2)}}{\sqrt{1 + \frac{\sigma_s^2 W}{\sigma_{y_{(2)}}^2} + \frac{\sigma_T^2(\sigma_{y_{(2)}}^2 + \mu_{y_{(2)}}^2)W}{\sigma_{y_{(2)}}^2}}}. \tag{IV.20}$$

Some existing mean estimators in the presence of measurement errors and non-response using the modified HH two-phase sampling are listed below:

- The ordinary mean estimator is given by

$$\hat{\mu}_{yw}^{HH} = \hat{\bar{y}}^* = w_1 \bar{y}_1^* + w_2 \hat{\bar{y}}_2^*. \tag{IV.21}$$

It can be written as

$$\hat{\mu}_{yw}^{HH} = (1 + e_0^*)\mu_y. \tag{IV.22}$$

The difference between the ordinary mean estimator and the true mean can be written as

$$\hat{\mu}_{yw}^{HH} - \mu_y = e_0^* \mu_y. \tag{IV.23}$$

Taking square and then expected value on both side of (IV.23), the MSE of $\hat{\mu}_{yw}$ is given by

$$MSE^*(\hat{\mu}_{yw}^{HH}) = \theta(\sigma_y^2 + \sigma_u^2) + \lambda(\sigma_{y_{(2)}}^2 + \sigma_p^2) + G \tag{IV.24}$$

where $G = \frac{W_2 f}{n}[\frac{[(\sigma_{y}{}^2_{(2)} + \mu_{y}{}^2_{(2)})\sigma_T^2 + \sigma_s^2]W}{(\mu_T W + 1 - W)^2}]$.

- A ratio estimator corresponding to the one in Gupta et al. (2014) is given by

$$\hat{\mu}_{rw}^{HH} = \frac{\hat{\bar{y}}^*}{\bar{x}^*}\mu_x = \hat{R}_w^{*HH}\mu_x, \tag{IV.25}$$

where $\hat{\bar{y}}^* = w_1\bar{y}_1^* + w_2\hat{\bar{y}}_2^*$.

It can be written as

$$\begin{aligned}\hat{\mu}_{rw}^{HH} &= \frac{(1+e_0^*)\mu_y}{(1+e_1^*)\mu_x}\mu_x \\ &= \mu_y(1+e_0^*)(1+e_1^*)^{-1} \\ &= \mu_y(1+e_0^*)(1-e_1^*+e_1^{*2}-e_1^{*3}+...) \\ &= \mu_y(1-e_1^*+e_1^{*2}+e_0^*-e_0^*e_1^*+...).\end{aligned} \tag{IV.26}$$

Using second order approximation, the difference between the ratio estimator and the true mean can be written as

$$\hat{\mu}_{yw}^{HH} - \mu_y = \mu_y(-e_1^* + e_1^{*2} + e_0^* - e_0^*e_1^*). \tag{IV.27}$$

Taking square and then expected value on both side of (IV.27), the MSE of $\hat{\mu}_{rw}^{HH}$ is given by

$$\begin{aligned}MSE^*(\hat{\mu}_{rw}^{HH}) = \theta(\sigma_y^2 + R^2\sigma_x^2 - 2R\rho_{yx}\sigma_y\sigma_x) + \lambda(\sigma_{y}{}^2_{(2)} + R^2\sigma_{x}{}^2_{(2)} - 2R\rho_{zx_{(2)}}\sigma_z \\ \sigma_{x(2)}) + \theta(\sigma_u^2 + R^2\sigma_v^2) + \lambda(\sigma_p^2 + R^2\sigma_v^2) + G,\end{aligned} \tag{IV.28}$$

where R$=\frac{\mu_y}{\mu_x}$.

The MSEs of the above mean estimators without measurement error may be obtained by letting $\sigma_u^2 = \sigma_p^2 = \sigma_v^2 = 0$ in (IV.24) and (IV.28).

IV.3. Generalized Estimator under ORRT Models and HH Two-phase Sampling Technique in the Presence of Measurement Errors

With this background, we use the generalized mean estimator used in Khalil et al. (2018) [34] and Chapter III. This estimator includes a wide variety of mean estimators as special cases. The non-response version is given by:

$$\hat{\mu}_{pw}^{HH} = (\hat{\bar{y}}^{*} + k(\mu_x - \bar{x}^{*}))(\frac{\bar{D}}{\bar{d}})^{v}, \tag{IV.29}$$

where $\hat{\bar{y}}^{*} = w_1\bar{y}_1^{*} + w_2\hat{\bar{y}}_2^{*}$, $\bar{x}^{*} = w_1\bar{x}_1^{*} + w_2\bar{x}_2^{*}$, $\bar{d} = \phi(\alpha\bar{x}^{*} + \beta) + (1-\phi)(\alpha\mu_x + \beta)$, $\bar{D} = \alpha\mu_x + \beta$, k and v are suitable constants. ϕ is assumed to be an unknown constant whose value is to be determined from optimally considerations. α ($\alpha \neq 0$) and β are assumed to be some known parameters of the auxiliary variable X, such as coefficient of variation (C_x), kurtosis, and correlation coefficient (ρ_{yx}) etc. Please note here with different values of α and β, we can obtain various estimators. Also, with v=1 we get various ratio estimators and with v=-1 we get various product estimators.

IV.3.1 Bias and MSE of the Generalized Mean Estimator

The generalized mean estimator will be studied under both modified HH two-phase sampling and measurement errors. According to the notations in Section IV.2, it can be written as

$$\hat{\mu}_{pw}^{HH} = ((1+e_0^{*})\mu_y + k(\mu_x - (1+e_1^{*})\mu_x))(\frac{\alpha\mu_x + \beta}{\phi(\alpha(1+e_1^{*})\mu_x + \beta) + (1-\phi)(\alpha\mu_x + \beta)})^{v}. \tag{IV.30}$$

Using Taylor's approximation and retaining terms of order up to 2, the difference between the generalized mean estimator and the true mean can be written as

$$
\begin{aligned}
&\hat{\mu}_{pw}^{HH} - \mu_y \\
&= ((1+e_0^*)\mu_y + k(\mu_x - (1+e_1^*)\mu_x))(\frac{\alpha\mu_x + \beta}{\phi(\alpha(1+e_1^*)\mu_x + \beta) + (1-\phi)(\alpha\mu_x + \beta)})^v - \mu_y \\
&= e_0^*\mu_y + e_1^*[-k\mu_x + \mu_y v(\frac{-\alpha\phi\mu_x}{\alpha\mu_x + \beta})] + \frac{1}{2!}[2e_0^*e_1^*\mu_y v(\frac{-\alpha\phi\mu_x}{\alpha\mu_x + \beta}) + e_1^{*2}(2kv\mu_x(\frac{\alpha\phi\mu_x}{\alpha\mu_x + \beta}) + \\
&\quad v(v+1)\mu_y(\frac{-\alpha\phi\mu_x}{\alpha\mu_x + \beta})^2)].
\end{aligned}
\tag{IV.31}
$$

Taking expectation on both side of (IV.31), the bias of the generalized mean estimator $\hat{\mu}_{pw}^{HH}$, correct to second order or approximation, is given by

$$
\begin{aligned}
Bias^*(\hat{\mu}_{pw}^{HH}) \approx &\theta[(kH + \frac{v+1}{v}\mu_y H^2)(\sigma_x^2 + \sigma_v^2) - H\rho_{yx}\sigma_y\sigma_x] + \\
&\lambda[(kH + \frac{v+1}{v}\mu_y H^2)(\sigma_{x(2)}^2 + \sigma_v^2) - H\rho_{zx(2)}\sigma_z\sigma_{x(2)}],
\end{aligned}
\tag{IV.32}
$$

where $H = \frac{\alpha\phi v}{\alpha\mu_x + \beta}$. The bias of $\hat{\mu}_{pw}^{HH}$ without measurement error may be obtained by setting $\sigma_v^2 = 0$ in above equation.

To determine the expression for MSE of the generalized mean estimator (IV.29), we take square of (IV.31) on both sides and retaining terms of order up to 2 which is given by

$$
\begin{aligned}
(\hat{\mu}_{pw}^{HH} - \mu_y)^2 = &e_0^{*2}\mu_y^2 + k^2\mu_x^2 e_1^{*2} + (H\mu_x\mu_y e_1^*)^2 - 2e_0^*e_1^*k\mu_x\mu_y - \\
&2e_0^*e_1^*H\mu_x\mu_y^2 + 2e_1^{*2}kH\mu_x^2\mu_y.
\end{aligned}
\tag{IV.33}
$$

Taking the expected value on both side of (IV.33), the expression for MSE of $\hat{\mu}_{pw}^{HH}$, correct to the first order approximation is given by

$$\begin{aligned}
MSE^*(\hat{\mu}_{pw}^{HH}) &= E(\hat{\mu}_{pw}^{HH} - \mu_y)^2 \\
&\cong \theta(\sigma_y^2 + k^2\sigma_x^2 + \phi^2 v^2 R_{pw}^2 \sigma_x^2 + 2k\phi v R_{pw}\sigma_x^2 - 2k\rho_{yx}\sigma_x\sigma_y - 2\phi v R_{pw}\rho_{yx} \\
&\quad \sigma_x\sigma_y) + \lambda(\sigma_{y_{(2)}}^2 + k^2\sigma_{x_{(2)}}^2 + \phi^2 v^2 R_{pw}^2 \sigma_{x_{(2)}}^2 + 2k\phi v R_{pw}\sigma_{x_{(2)}}^2 - 2k\rho_{zx_{(2)}} \\
&\quad \sigma_z\sigma_{x_{(2)}} - 2\phi v R_{pw}\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}) + \theta(\sigma_u^2 + k^2\sigma_v^2 + \phi^2 v^2 R_{pw}^2\sigma_v^2 + \\
&\quad 2k\phi v R_{pw}\sigma_v^2) + \lambda(\sigma_p^2 + k^2\sigma_v^2 + \phi^2 v^2 R_{pw}^2\sigma_v^2 + 2k\phi v R_{pw}\sigma_v^2) + G \\
&= \theta[\sigma_y^2 + (k + \phi v R_{pw})^2\sigma_x^2 - 2(k + \phi v R_{pw})\rho_{yx}\sigma_x\sigma_y] + \\
&\quad \lambda[\sigma_{y(2)}^2 + (k + \phi v R_{pw})^2\sigma_{x(2)}^2 - 2(k + \phi v R_{pw})\rho_{zx(2)}\sigma_x\sigma_z] + \\
&\quad \theta[\sigma_u^2 + (k + \phi v R_{pw})^2\sigma_v^2] + \lambda[\sigma_p^2 + (k + \phi v R_{pw})^2\sigma_v^2] + G,
\end{aligned} \tag{IV.34}$$

where $R_{pw} = \frac{\alpha\mu_y}{\alpha\mu_x + \beta}$. Minimization of the above expression (IV.34) with respect to ϕ yields its optimum value as:

$$\phi_{opt} \cong \frac{\theta(\rho_{yx}\sigma_x\sigma_y - k(\sigma_x^2 + \sigma_v^2)) + \lambda(\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}} - k(\sigma_{x_{(2)}}^2 + \sigma_v^2))}{vR_{pw}[\theta(\sigma_x^2 + \sigma_v^2) + \lambda(\sigma_{x_{(2)}}^2 + \sigma_v^2)]}. \tag{IV.35}$$

Substitution of ϕ_{opt} in MSE($\hat{\mu}_{pw}^{HH}$) yields the minimum value as:

$$\begin{aligned}
MSE_{min}^*(\hat{\mu}_{pw}^{HH}) \cong& \theta(\sigma_y^2 + P^2\sigma_x^2 - 2P\rho_{yx}\sigma_x\sigma_y) + \lambda(\sigma_{y_{(2)}}^2 + P^2\sigma_{x_{(2)}}^2 - 2P\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}) + \\
& \theta(\sigma_u^2 + P^2\sigma_v^2) + \lambda(\sigma_p^2 + P^2\sigma_v^2) + G,
\end{aligned} \tag{IV.36}$$

where $P = \frac{\theta\rho_{yx}\sigma_x\sigma_y + \lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}}{\theta(\sigma_x^2 + \sigma_v^2) + \lambda(\sigma_{x_{(2)}}^2 + \sigma_v^2)}$.

The expression for the minimized MSE of the generalized estimator without ME may be obtained by putting $\sigma_u^2 = \sigma_v^2 = \sigma_p^2 = 0$ in the (IV.36), which gives

$$\begin{aligned} MSE_{min}(\hat{\mu}_{pw}^{HH}) \cong& \theta(\sigma_y^2 + P^2\sigma_x^2 - 2P\rho_{yx}\sigma_x\sigma_y) + \lambda(\sigma_{y_{(2)}}^2 + P^2\sigma_{x_{(2)}}^2 - 2P\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}) \\ &+ G, \end{aligned} \tag{IV.37}$$

where $G = \frac{W_2 f}{n}[\frac{\sigma_s^2 W + \sigma_T^2 W(\sigma_{y_{(2)}}^2 + \mu_{y_{(2)}}^2)}{(\mu_T W + 1 - W)^2}]$.

IV.3.2 Efficiency Comparisons

Comparing the MSE expressions of $\hat{\mu}_{yw}^{HH}$ (IV.24), $\hat{\mu}_{rw}^{HH}$ (IV.28), and $\hat{\mu}_{pw}^{HH}$ (IV.36) with measurement errors, it can be verified easily that

- $MSE_{min}^*(\hat{\mu}_{pw}^{HH}) < MSE^*(\hat{\mu}_{yw}^{HH})$ if

$$\begin{aligned} MSE_{min}^*(\hat{\mu}_{pw}^{HH}) - MSE^*(\hat{\mu}_{yw}^{HH}) =& P^2(\theta(\sigma_x^2 + \sigma_v^2) + \lambda(\sigma_{x_{(2)}}^2 + \sigma_v^2)) \\ &- 2P(\theta\rho_{yx}\sigma_x\sigma_y + \lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}) \\ =& -\frac{(\theta\rho_{yx}\sigma_x\sigma_y + \lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}})^2}{\theta(\sigma_x^2 + \sigma_v^2) + \lambda(\sigma_{x_{(2)}}^2 + \sigma_v^2)} < 0, \end{aligned} \tag{IV.38}$$

- $MSE_{min}^*(\hat{\mu}_{pw}^{HH}) < MSE^*(\hat{\mu}_{rw}^{HH})$ if

$$\begin{aligned} MSE_{min}^*(\hat{\mu}_{pw}^{HH}) - MSE^*(\hat{\mu}_{rw}^{HH}) =& (P^2 - R^2)(\theta(\sigma_x^2 + \sigma_v^2) + \lambda(\sigma_{x_{(2)}}^2 + \sigma_v^2)) \\ &- 2(P + R)(\theta\rho_{yx}\sigma_x\sigma_y + \lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}^2) < 0; \end{aligned} \tag{IV.39}$$

In other words, if

$$\begin{aligned}\frac{(P^2-R^2)(\theta(\sigma_x^2+\sigma_v^2)+\lambda(\sigma_{x_{(2)}}^2+\sigma_v^2))}{2(P+R)(\theta\rho_{yx}\sigma_x\sigma_y+\lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}})} &= \frac{(P-R)(\theta(\sigma_x^2+\sigma_v^2)+\lambda(\sigma_{x_{(2)}}^2+\sigma_v^2))}{2(\theta\rho_{yx}\sigma_x\sigma_y+\lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}})}\\ &= \frac{P-R}{2P}\\ &= \frac{1}{2}-\frac{R}{2P}\\ &= \frac{1}{2}-\frac{\mu_y}{2\mu_x}\frac{\theta(\sigma_x^2+\sigma_v^2)+\lambda(\sigma_{x_{(2)}}^2+\sigma_v^2)}{\theta\rho_{yx}\sigma_x\sigma_y+\lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}} < 1;\end{aligned} \tag{IV.40}$$

and

- $MSE^*(\hat{\mu}_{rw}^{HH}) < MSE^*(\hat{\mu}_{yw}^{HH})$ if

$$\begin{aligned}MSE^*(\hat{\mu}_{rw}^{HH})-MSE^*(\hat{\mu}_{yw}^{HH}) =& R^2(\theta(\sigma_x^2+\sigma_v^2)+\lambda(\sigma_{x_{(2)}}^2+\sigma_v^2))\\ &-2R(\theta\rho_{yx}\sigma_x\sigma_y+\lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}) < 0;\end{aligned} \tag{IV.41}$$

In other words, if

$$\frac{R^2}{2R}\frac{\theta(\sigma_x^2+\sigma_v^2)+\lambda(\sigma_{x_{(2)}}^2+\sigma_v^2)}{\theta\rho_{yx}\sigma_x\sigma_y+\lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}} = \frac{\mu_y}{2\mu_x}\frac{\theta(\sigma_x^2+\sigma_v^2)+\lambda(\sigma_{x_{(2)}}^2+\sigma_v^2)}{\theta\rho_{yx}\sigma_x\sigma_y+\lambda\rho_{zx_{(2)}}\sigma_z\sigma_{x_{(2)}}} < 1. \tag{IV.42}$$

The conditions (IV.38) and (IV.40) always hold true. From (IV.42), the ratio estimator is more efficient than the ordinary mean estimator only if the measurement error on auxiliary variable X (σ_v^2) is small, and X and Y are highly correlated.

IV.4. Simulation Study

In this section, we will evaluate the performance of the generalized mean estimator under non-response and measurement errors with the other two estimators by a simulation study. In the generalized mean estimator, we choose v and k to be 1, and ϕ to be its optimum value. As demonstrated in Chapter III simulations, we could use various parameters associated with the auxiliary variable such as the coefficient of variation (C_x) or kurtosis for α and β, but these choices do not impact the results in any meaningful way. We will only show the results where $\alpha = 1$ and $\beta = 0$. The scrambling variable S is taken to be a normal variate with mean equal to zero and vary variance ($0.2{*}\sigma_x^2$, $0.5{*}\sigma_x^2$, and $1{*}\sigma_x^2$). And T is also taken to be a normal variate but with mean equal to one and varying variances (0, 0.5, 1). The measurement errors of X have a normal distribution with mean zero in both phases; the measurement errors of Y in the first phase and Z in the second phase have a normal distribution with mean zero. We demonstrate different variances (0, 5, 10) for measurement errors. We consider a finite population of size 5000 generated from bivariate normal distribution with means and covariance of (Y, X) as given below.

$$\textbf{Population} \quad \mu = \begin{bmatrix} 10 \\ 6 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 16 & 9.051 \\ 9.051 & 8 \end{bmatrix}, \quad \rho_{yx} = 0.8$$

The real parameters of the set of 5000 data points we generated using R are very close to the parameter values in (A) but not exactly same. For the simulation study, we used parameter values in (B) and not those in (A).

$$\mu_x = 6,\ \sigma_x^2 = 8,\ \mu_y = 10,\ \sigma_y^2 = 16,\ \rho_{yx} = 0.8 \tag{A}$$

$$\mu_x = 6.0228,\ \sigma_x^2 = 8.1830,\ \mu_y = 9.9864,\ \sigma_y^2 = 16.1215,\ \rho_{yx} = 0.8024 \tag{B}$$

We consider samples of size $n = 500$ using SRSWOR (simple random sampling without replacement) and assume response rate is 40% in the first phase. This means in the first phase only 200 (n_1) subjects provide a response to the survey question and 300 (n_2) of them do not. In the second phase, we take another sample ($n_s = \frac{n_2}{f}$) from non-respondent group by using f=2, 3, 4, respectively. Different response rates of 20%, 40% and 60% are also compared in the simulation study. Coding for the simulations was done in R and results are averaged over 5,000 iterations. The empirical MSE of the estimator $\hat{\mu}_y$ is computed by

$$MSE^*(\hat{\mu}_w) = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\mu}_w^{HH} - \mu_y)^2, \tag{IV.43}$$

where $\hat{\mu}_w^{HH} = \hat{\mu}_{yw}^{HH}$, $\hat{\mu}_{rw}^{HH}$, and $\hat{\mu}_{pw}^{HH}$. Here, μ_y is the population mean of the sensitive study variable.

The percent relative efficiencies (PREs) of the estimators ($\hat{\mu}_w^{HH}$) with respect to the ordinary mean estimator ($\hat{\mu}_{yw}^{HH}$) is defined as

$$PRE = \frac{MSE^*(\hat{\mu}_{yw}^{HH})}{MSE^*(\hat{\mu}_w^{HH})} * 100. \tag{IV.44}$$

We will also use the unified measure δ of efficiency and privacy as defined in Gupta et al. (2018)[24]. It is given by

$$\delta = \frac{MSE^*(\hat{\mu}_w^{HH})}{\Delta_{DP}}. \tag{IV.45}$$

In (IV.45), MSE is used in place of Var(.) to account for biased estimators.

Table IV.1. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators when Response Rate = 40% , σ_v^2 = σ_u^2= σ_p^2 = 1, f = 2 and σ_s^2 = 0.2*σ_x^2.

Est.	W	σ_T^2	MSE		PRE		δ
			Without ME	With ME	Without ME	With ME	
$\hat{\mu}_{yw}^{HH}$	0.5	0	**0.0508**	**0.0537**	**100.0000**	**100.0000**	**0.0330**
			0.0514	0.0551	100.0000	100.0000	0.0339
		0.5	**0.1190**	**0.1220**	**100.0000**	**100.0000**	**0.0021**
			0.1140	0.1164	100.0000	100.0000	0.0020
		1	**0.1865**	**0.1895**	**100.0000**	**100.0000**	**0.0016**
			0.1787	0.1810	100.0000	100.0000	0.0016
	0.8	0	**0.0519**	**0.0549**	**100.0000**	**100.0000**	**0.0337**
			0.0517	0.0554	100.0000	100.0000	0.0340
		0.5	**0.1603**	**0.1633**	**100.0000**	**100.0000**	**0.0028**
			0.1603	0.1631	100.0000	100.0000	0.0028
		1	**0.2670**	**0.2700**	**100.0000**	**100.0000**	**0.0023**
			0.2698	0.2723	100.0000	100.0000	0.0023
	1	0	**0.0527**	**0.0557**	**100.0000**	**100.0000**	**0.0342**
			0.0529	0.0567	100.0000	100.0000	0.0348
		0.5	**0.1875**	**0.1905**	**100.0000**	**100.0000**	**0.0032**
			0.1790	0.1851	100.0000	100.0000	0.0031
		1	**0.3197**	**0.3227**	**100.0000**	**100.0000**	**0.0028**
			0.3169	0.3158	100.0000	100.0000	0.0027
$\hat{\mu}_{rw}^{HH}$	0.5	0	**0.0261**	**0.0373**	**194.6360**	**143.9678**	**0.0229**
			0.0269	0.0388	191.0781	142.0103	0.0238
		0.5	**0.0943**	**0.1055**	**126.1930**	**115.6398**	**0.0018**
			0.0910	0.1055	125.2747	110.3318	0.0018
		1	**0.1618**	**0.1730**	**115.2658**	**109.5376**	**0.0015**
			0.1566	0.1650	114.1124	109.6970	0.0014
	0.8	0	**0.0273**	**0.0385**	**190.1099**	**142.5974**	**0.0237**
			0.0283	0.0385	182.6855	143.8961	0.0237
		0.5	**0.1356**	**0.1468**	**118.2153**	**111.2398**	**0.0025**
			0.1381	0.1453	116.0753	112.2505	0.0025

		1	**0.2423**	**0.2534**	**110.1940**	**106.5509**	**0.0022**
			0.2486	0.2542	108.5278	107.1204	0.0022
	1	0	**0.0280**	**0.0393**	**188.2143**	**141.7303**	**0.0242**
			0.0284	0.0391	186.2676	145.0128	0.0240
		0.5	**0.1628**	**0.1740**	**115.1720**	**109.4828**	**0.0029**
			0.1588	0.1681	112.7204	110.1130	0.0028
		1	**0.2950**	**0.3062**	**108.3729**	**105.3886**	**0.0026**
			0.2888	0.2994	109.7299	105.4776	0.0026
$\hat{\mu}_{pw}^{HH}$	0.5	0	**0.0191**	**0.0255**	**265.9686**	**210.5882**	**0.0157**
			0.0196	0.0265	262.2449	207.9245	0.0163
		0.5	**0.0873**	**0.0937**	**136.3116**	**130.2028**	**0.0016**
			0.0835	0.0985	136.5269	118.1726	0.0017
		1	**0.1549**	**0.1612**	**120.4003**	**117.5558**	**0.0014**
			0.1488	0.1634	120.0941	110.7711	0.0014
	0.8	0	**0.0203**	**0.0267**	**255.6650**	**205.6180**	**0.0164**
			0.0206	0.0276	250.9709	200.7246	0.0170
		0.5	**0.1286**	**0.1350**	**124.6501**	**120.9630**	**0.0023**
			0.1303	0.1347	123.0238	121.0839	0.0023
		1	**0.2353**	**0.2417**	**113.4722**	**111.7087**	**0.0021**
			0.2405	0.2440	112.1830	111.5984	0.0021
	1	0	**0.0211**	**0.0274**	**249.7630**	**203.2847**	**0.0168**
			0.0216	0.0280	244.9074	202.5000	0.0172
		0.5	**0.1559**	**0.1622**	**120.2694**	**117.4476**	**0.0027**
			0.1499	0.1565	119.4129	118.2748	0.0026
		1	**0.2880**	**0.2944**	**111.0069**	**109.6128**	**0.0025**
			0.2793	0.2879	113.4622	109.6909	0.0025

Table IV.2. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators when Response Rate = 40% , $\sigma_v^2 = \sigma_u^2 = \sigma_p^2 = 1$, f = 2 and $\sigma_s^2 = 0.5{*}\sigma_x^2$.

Est.	W	σ_T^2	MSE Without ME	MSE With ME	PRE Without ME	PRE With ME	δ
$\hat{\mu}_{yw}^{HH}$	0.5	0	**0.0537**	**0.0567**	**100.0000**	**100.0000**	**0.0139**
			0.0545	0.0584	100.0000	100.0000	0.0143
		0.5	**0.1219**	**0.1249**	**100.0000**	**100.0000**	**0.0020**
			0.1172	0.1196	100.0000	100.0000	0.0020
		1	**0.1895**	**0.1924**	**100.0000**	**100.0000**	**0.0016**
			0.1820	0.1843	100.0000	100.0000	0.0016
	0.8	0	**0.0566**	**0.0596**	**100.0000**	**100.0000**	**0.0146**
			0.0559	0.0598	100.0000	100.0000	0.0146
		0.5	**0.1650**	**0.1680**	**100.0000**	**100.0000**	**0.0027**
			0.1651	0.1679	100.0000	100.0000	0.0027
		1	**0.2716**	**0.2746**	**100.0000**	**100.0000**	**0.0023**
			0.2746	0.2771	100.0000	100.0000	0.0023
	1	0	**0.0586**	**0.0616**	**100.0000**	**100.0000**	**0.0151**
			0.0587	0.0625	100.0000	100.0000	0.0153
		0.5	**0.1933**	**0.1963**	**100.0000**	**100.0000**	**0.0032**
			0.1847	0.1916	100.0000	100.0000	0.0031
		1	**0.3255**	**0.3285**	**100.0000**	**100.0000**	**0.0028**
			0.3225	0.3226	100.0000	100.0000	0.0027
$\hat{\mu}_{rw}^{HH}$	0.5	0	**0.0290**	**0.0403**	**185.1724**	**140.6948**	**0.0099**
			0.0301	0.0420	181.0631	139.0476	0.0103
		0.5	**0.0972**	**0.1084**	**125.4115**	**115.2214**	**0.0018**
			0.0940	0.1028	124.6809	116.3424	0.0017
		1	**0.1648**	**0.1759**	**114.9879**	**109.3803**	**0.0015**
			0.1597	0.1679	113.9637	109.7677	0.0014
	0.8	0	**0.0319**	**0.0432**	**177.4295**	**137.9630**	**0.0106**
			0.0330	0.0454	169.3939	131.7181	0.0111
		0.5	**0.1403**	**0.1514**	**117.6051**	**110.9643**	**0.0025**
			0.1425	0.1498	115.8596	112.0828	0.0024

		1	**0.2469**	**0.2581**	**110.0041**	**106.3929**	**0.0022**
			0.2531	0.2588	108.4947	107.0711	0.0022
	1	0	**0.0339**	**0.0451**	**172.8614**	**136.5854**	**0.0110**
			0.0345	0.0451	170.1449	138.5809	0.0110
		0.5	**0.1686**	**0.1798**	**114.6501**	**109.1769**	**0.0029**
			0.1640	0.1741	112.6220	110.0517	0.0028
		1	**0.3008**	**0.3119**	**108.2114**	**105.3222**	**0.0026**
			0.2940	0.3058	109.6939	105.4938	0.0026
$\hat{\mu}_{pw}^{HH}$	0.5	0	**0.0220**	**0.0284**	**244.0909**	**199.6479**	**0.0069**
			0.0228	0.0296	239.0351	197.2973	0.0072
		0.5	**0.0903**	**0.0966**	**134.9945**	**129.2961**	**0.0016**
			0.0864	0.0912	135.6481	131.1404	0.0015
		1	**0.1578**	**0.1641**	**120.0887**	**117.2456**	**0.0014**
			0.1518	0.1562	119.8946	117.9898	0.0013
	0.8	0	**0.0250**	**0.0313**	**226.4000**	**190.4153**	**0.0077**
			0.0252	0.0322	221.8254	185.7143	0.0079
		0.5	**0.1333**	**0.1397**	**123.7809**	**120.2577**	**0.0023**
			0.1348	0.1392	122.4777	120.6178	0.0023
		1	**0.2399**	**0.2463**	**113.2138**	**111.4901**	**0.0021**
			0.2451	0.2485	112.0359	111.5091	0.0021
	1	0	**0.0269**	**0.0333**	**217.8439**	**184.9850**	**0.0081**
			0.0276	0.0339	212.6812	184.3658	0.0083
		0.5	**0.1617**	**0.1680**	**119.5424**	**116.8452**	**0.0027**
			0.1553	0.1629	118.9311	117.6182	0.0027
		1	**0.2938**	**0.3002**	**110.7897**	**109.4270**	**0.0025**
			0.2847	0.2944	113.2771	109.5788	0.0025

Table IV.3. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators when Response Rate = 40% , $\sigma_v^2 = \sigma_u^2 = \sigma_p^2 = 1$, f = 2 and $\sigma_s^2 = 1{*}\sigma_x^2$.

Est.	W	σ_T^2	MSE Without ME	MSE With ME	PRE Without ME	PRE With ME	δ
$\hat{\mu}_{yw}^{HH}$	0.5	0	**0.0586**	**0.0616**	**100.0000**	**100.0000**	**0.0076**
			0.0597	0.0637	100.0000	100.0000	0.0078
		0.5	**0.1268**	**0.1298**	**100.0000**	**100.0000**	**0.0020**
			0.1224	0.1247	100.0000	100.0000	0.0019
		1	**0.1943**	**0.1973**	**100.0000**	**100.0000**	**0.0016**
			0.1873	0.1895	100.0000	100.0000	0.0015
	0.8	0	**0.0644**	**0.0674**	**100.0000**	**100.0000**	**0.0083**
			0.0632	0.0672	100.0000	100.0000	0.0083
		0.5	**0.1728**	**0.1757**	**100.0000**	**100.0000**	**0.0027**
			0.1728	0.1757	100.0000	100.0000	0.0027
		1	**0.2793**	**0.2823**	**100.0000**	**100.0000**	**0.0023**
			0.2824	0.2849	100.0000	100.0000	0.0023
	1	0	**0.0683**	**0.0713**	**100.0000**	**100.0000**	**0.0088**
			0.0685	0.0722	100.0000	100.0000	0.0089
		0.5	**0.2030**	**0.2060**	**100.0000**	**100.0000**	**0.0031**
			0.2042	0.2021	100.0000	100.0000	0.0031
		1	**0.3351**	**0.3380**	**100.0000**	**100.0000**	**0.0027**
			0.3320	0.3334	100.0000	100.0000	0.0027
$\hat{\mu}_{rw}^{HH}$	0.5	0	**0.0339**	**0.0451**	**172.8614**	**136.5854**	**0.0055**
			0.0353	0.0471	169.1218	135.2442	0.0058
		0.5	**0.1021**	**0.1133**	**124.1920**	**114.5631**	**0.0017**
			0.0999	0.1176	122.5225	106.0374	0.0018
		1	**0.1696**	**0.1808**	**114.5637**	**109.1261**	**0.0015**
			0.1648	0.1729	113.6529	109.6009	0.0014
	0.8	0	**0.0398**	**0.0510**	**161.8090**	**132.1569**	**0.0063**
			0.0410	0.0534	154.1463	125.8427	0.0066
		0.5	**0.1481**	**0.1592**	**116.6779**	**110.3643**	**0.0024**
			0.1499	0.1573	115.2769	111.6974	0.0024

		1	**0.2546**	**0.2658**	**109.7015**	**106.2077**	**0.0022**
			0.2606	0.2663	108.3653	106.9846	0.0022
		0	**0.0437**	**0.0549**	**156.2929**	**129.8725**	**0.0067**
			0.0445	0.0550	153.9326	131.2727	0.0068
	1	0.5	**0.1783**	**0.1895**	**113.8531**	**108.7071**	**0.0029**
			0.1730	0.1842	118.0347	109.7177	0.0028
		1	**0.3104**	**0.3215**	**107.9575**	**105.1322**	**0.0026**
			0.3129	0.3160	106.1042	105.5063	0.0026
		0	**0.0269**	**0.0333**	**217.8439**	**184.9850**	**0.0041**
			0.0279	0.0348	213.9785	183.0460	0.0043
	0.5	0.5	**0.0952**	**0.1015**	**133.1933**	**127.8818**	**0.0015**
			0.0916	0.0964	133.6245	129.3568	0.0015
		1	**0.1627**	**0.1690**	**119.4222**	**116.7456**	**0.0014**
			0.1571	0.1616	119.2234	117.2649	0.0013
		0	**0.0328**	**0.0392**	**196.3415**	**171.9388**	**0.0048**
			0.0328	0.0399	192.6829	168.4211	0.0049
$\hat{\mu}_{pw}^{HH}$	0.8	0.5	**0.1411**	**0.1474**	**122.4663**	**119.1995**	**0.0022**
			0.1423	0.1469	121.4336	119.6052	0.0022
		1	**0.2477**	**0.2540**	**112.7574**	**111.1417**	**0.0021**
			0.2477	0.2562	114.0089	111.2022	0.0021
		0	**0.0367**	**0.0431**	**186.1035**	**165.4292**	**0.0053**
			0.0375	0.0436	182.6667	165.5963	0.0054
	1	0.5	**0.1713**	**0.1777**	**118.5055**	**115.9257**	**0.0027**
			0.1744	0.1731	117.0872	116.7533	0.0026
		1	**0.3034**	**0.3097**	**110.4483**	**109.1379**	**0.0025**
			0.3037	0.3049	109.3184	109.3473	0.0025

Tables IV.1, IV.2, IV.3 present the theoretical and empirical MSEs and PREs of the ORRT mean estimators when all the variances of measurement errors (σ_v^2, σ_u^2 and σ_p^2) are set equal to 1 and response rate in Phase I is set equal to 40% with different variances of S ($0.2^*\sigma_x^2$, $0.5^*\sigma_x^2$, $1^*\sigma_x^2$), respectively. Comparing these three tables, the mean estimation is less efficient as the variance of S increases in the presence of non-response. For example, under the situation when variance of T is equal to 0.5, the sensitivity level W is equal to 0.8, and in the presence of measurement errors, the MSEs of the generalized mean estimator are respectively equal to 0.1350, 0.1397 and 0.1774 for the variance of S equal to 0.2*σ_x^2, 0.5*σ_x^2, and σ_x^2. These results are consistent with the theoretical results. Larger variance of S introduces larger penalty for using RRT models.

For all three tables, the MSE of the mean estimators increases as W increases under non-response and measurement errors. For example, in Table IV.1, the MSE of the generalized mean estimator increased from 0.0937 to 0.1622 as the sensitivity level increased from 0.5 to 1 when variance of T is equal to 0.5. It indicates that the ORRT model gains some efficiency when some of the respondents feel the survey question is not sensitive. Furthermore, as the variance of T increases, the MSE increases while δ decreases with a reasonably small value of σ_T^2. For instance, in Table IV.1, when the sensitivity level W is equal to 0.5, the MSE of the generalized mean estimator increases from 0.0333 to 0.1690 as the variance of T increases from 0 to 1, while the δ value decreases from 0.0041 to 0.0014. Similar to Chapter III, mean estimators under the general linear combination model (Z=TY+S) is better than under the simple additive model (Z=Y+S) when non-response is present if both efficiency and privacy are considered through the unified measures.

Table IV.4. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators under the Conditions of σ_v^2 = σ_u^2= σ_p^2 = 1, 5, 10 when Response Rate = 40% , and σ_s^2 = 0.5*σ_x^2.

Est.	f	MSE			PRE		
		1	5	10	1	5	10
$\hat{\mu}_{yw}^{HH}$	2	**0.1680**	**0.1799**	**0.1948**	**100.0000**	**100.0000**	**100.0000**
		0.1679	0.1799	0.1951	100.0000	100.0000	100.0000
	3	**0.2470**	**0.2637**	**0.2847**	**100.0000**	**100.0000**	**100.0000**
		0.2397	0.2653	0.2854	100.0000	100.0000	100.0000
	4	**0.3261**	**0.3476**	**0.3745**	**100.0000**	**100.0000**	**100.0000**
		0.3173	0.3484	0.3756	100.0000	100.0000	100.0000
$\hat{\mu}_{rw}^{HH}$	2	**0.1514**	**0.1960**	**0.2518**	**110.9643**	**91.7857**	**77.3630**
		0.1498	0.1894	0.2407	112.0828	94.9842	81.0553
	3	**0.2236**	**0.2859**	**0.3638**	110.4651	**92.2350**	**78.2573**
		0.2223	0.2882	0.3611	107.8273	92.0541	79.0363
	4	**0.2957**	**0.3758**	**0.4758**	**110.2807**	**92.4960**	**78.7095**
		0.2854	0.3761	0.4705	111.1773	92.6349	79.8300
$\hat{\mu}_{pw}^{HH}$	2	**0.1397**	**0.1601**	**0.1804**	**120.2577**	**112.3673**	**107.9823**
		0.1392	0.1585	0.1790	120.6178	113.5016	108.9944
	3	**0.2071**	**0.2357**	**0.2642**	**119.2661**	**111.8795**	**107.7593**
		0.1972	0.2353	0.2626	121.5517	112.7497	108.6824
	4	**0.2745**	**0.3113**	**0.3480**	**118.7978**	**111.6608**	**107.6149**
		0.2653	0.3114	0.3486	119.6005	111.8818	107.7453

Table IV.4 presents the theoretical and empirical MSEs and PREs of the ORRT mean estimators under different variances of measurement errors (1, 5, 10) when the sensitivity level W is 0.8, variance of T is 0.5 and response rate in Phase I is 40%. As the measurement errors increase, the MSE of each mean estimator increases. For instance, the MSE of the generalized mean estimator increased from 0.1297 to 0.1894 as the variance of measurement errors increased from 1 to 10 when the value of f is

2. It is obvious that larger measurement errors have larger negative impact on mean estimation under non-response.

Also, from Tables IV.1, IV.2, IV.3 and Table IV.4, it is clear that the generalized mean estimator $\hat{\mu}_{pw}$ is more efficient than the other two mean estimators even when very large measurement errors are present. However, the ratio estimator $\hat{\mu}_{rw}$ becomes less efficient than the ordinary mean estimator $\hat{\mu}_{yw}$ as the measurement errors increase. For example in Table IV.4, the MSE of the generalized mean estimator 0.1804 is less than the MSE of the ordinary mean estimator 0.1948 when the variance of measurement errors is 10. However, the MSE of the ratio estimator 0.2518 is larger than other two estimators because the measurement errors are large. The reason is the measurement errors take place on both X and Y for the ratio estimator and only on Y for the ordinary mean estimator. This result shows the superiority of the generalized mean estimator in the presence of measurement errors and non-response because it is not affected as badly as the ratio estimator by measurement errors on X.

Table IV.5 presents the theoretical and empirical MSEs and PREs of the ORRT mean estimators under different response rates when the variance of measurement errors is equal to 1, sensitivity level W is equal to 0.8, and variance of T is equal to 0.5. The efficiency of each estimator gets better as the response rate increases. In other words, the larger the sample we collect from the first call, the higher is the efficiency of the mean estimation.

In addition, from both Tables IV.4 and IV.5, the efficiency of each estimator gets worse as the value of f increases. For example, the MSE of the generalized mean estimator increased from 0.1601 to 0.3113 as the value of f increased from 2 to 4 when the variance of measurement errors is 5. It is reasonable because larger f value means we obtain smaller sample from the second call.

Table IV.5. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators under the Conditions of Response Rate (RR) = 20%, 40%, 60% when $\sigma_v^2 = \sigma_u^2 = \sigma_p^2 = 1$, W = 0.8, and $\sigma_T^2 = 0.5*\sigma_x^2$

RR.	f	MSE			PRE		
		20%	40%	60%	20%	40%	60%
$\hat{\mu}_{yw}^{HH}$	2	**0.2134**	**0.1680**	**0.1231**	**100.0000**	**100.0000**	**100.0000**
		0.2137	0.1679	0.1231	100.0000	100.0000	100.0000
	3	**0.3184**	**0.2470**	**0.1764**	**100.0000**	**100.0000**	**100.0000**
		0.3110	0.2397	0.1775	100.0000	100.0000	100.0000
	4	**0.4234**	**0.3261**	**0.2298**	**100.0000**	**100.0000**	**100.0000**
		0.4242	0.3173	0.2311	100.0000	100.0000	100.0000
$\hat{\mu}_{rw}^{HH}$	2	**0.1951**	**0.1514**	**0.1086**	**109.3798**	**110.9643**	**113.3517**
		0.1953	0.1498	0.1042	109.4214	112.0828	118.1382
	3	**0.2914**	**0.2236**	**0.1570**	**109.2656**	**110.4651**	**112.3567**
		0.2930	0.2223	0.1548	106.1433	107.8273	114.6641
	4	**0.3878**	**0.2957**	**0.2055**	**109.1800**	**110.2807**	**111.8248**
		0.3874	0.2854	0.2001	109.4992	111.1773	115.4923
$\hat{\mu}_{pw}^{HH}$	2	**0.1818**	**0.1397**	**0.0983**	**117.3817**	**120.2577**	**125.2289**
		0.1826	0.1392	0.0954	117.0318	120.6178	129.0356
	3	**0.2718**	**0.2071**	**0.1436**	**117.1450**	**119.2661**	**122.8412**
		0.2731	0.1972	0.1428	113.8777	121.5517	124.2997
	4	**0.3618**	**0.2745**	**0.1888**	**117.0260**	**118.7978**	**121.7161**
		0.3585	0.2653	0.1870	118.3264	119.6005	123.5829

IV.5. Concluding Chapter Remarks

The main contribution in this chapter is the mean estimation of a sensitive variable, in the presence of measurement errors and non-response, using modified HH two-phase sampling. ORRT model leads to better results than non-optional RRT model under the presence of non-response and measurement errors simultaneously. Measurement errors have a negative impact on mean estimation, especially when they are large. A simulation study verifies the theoretical results. It is also clear from the theoretical conditions (IV.38), (IV.40), (IV.42) and the simulation results that the generalized mean estimator is always more efficient that the ordinary RRT mean estimator and the ratio estimator, while the ratio estimator is less efficient than the ordinary mean estimator if the measurement errors on X are large.

We have used SRS in the current study. We want to explore the performance of mean estimation under the same conditions as in this Chapter, but using stratified random sampling. The mean estimation of a sensitive variable under both measurement errors and non-response using stratified random sampling will be discussed in the next chapter.

CHAPTER V

MEAN ESTIMATION IN THE SIMULTANEOUS PRESENCE OF MEASUREMENT ERRORS AND NON-RESPONSE USING ORRT MODELS UNDER THE STRATIFIED RANDOM SAMPLING DESIGN

We have so far discussed mean estimation under measurement errors, and/or non-response using simple random sampling[92]. In this Chapter, we will continue the Chapter IV work but using stratified random sampling. In Section V.1, some existing mean estimators under measurement errors and non-response will be presented; in Section V.2, the generalized mean estimator will be discussed; Section V.3 will present a simulation study; and Section V.4 will provide concluding remarks for this Chapter.

V.1. Some Existing Mean Estimators under Measurement Errors and Non-response using Stratified Random Sampling

Let a finite population $U = (U_1, U_2, U_3, .., U_N)$ is divided in L homogeneous strata with N_h representing the number of units in stratum h such that $\sum_{h=1}^{L} N_h = N$. From h^{th} stratum, a simple random sample of size n_h is drawn without replacement such that $\sum_{h=1}^{L} n_h = n$. Under the situation where non-response is present, we assume that n_{1h} units provided response on the first call and remaining $n_{2h} = n_h - n_{1h}$ units do not respond. Then a sub-sample of size $n_{sh} = \frac{n_{2h}}{f_h}$ $(f_h > 1)$ is taken from the n_{2h} non-response units in the h^{th} stratum. We use standard terminology, as used in Chapter IV, but with term 'h' (h = 1, 2, ..., h) in the subscript to represent terms in the h^{th} stratum. In the h^{th} stratum, let the measurement errors of the auxiliary variable (X) be given by $V_{ih} = x_{ih} - X_{ih}$. Let the respective measurement errors associated with the study variable (Y) in the population and the scrambled variable (Z) in the face-to-face phase be given by $U_{ih} = y_{ih} - Y_{ih}$ and $P_{ih} = z_{ih} - Z_{ih}$. These measurement errors are assumed to be random and uncorrelated with mean zero and variances σ^2_{vh}, σ^2_{uh}, and σ^2_{ph}, respectively.

Assume population mean of the auxiliary variable μ_x is known, and non-response happens on both X and Y. Some notations are given below

$$\Omega_y = \sum_{h=1}^{L}\sum_{i=1}^{n_h}(y_{ih} - \mu_{yh}), \tag{V.1}$$

$$\Omega_x = \sum_{h=1}^{L}\sum_{i=1}^{n_h}(x_{ih} - \mu_{xh}), \tag{V.2}$$

$$\Omega_u = \sum_{h=1}^{L}(\sum_{i=1}^{n_{1h}}U_{ih} + \sum_{i=1}^{n_{2h}}P_{ih}), \tag{V.3}$$

and

$$\Omega_v = \sum_{h=1}^{L}\sum_{i=1}^{n_h}V_{ih}. \tag{V.4}$$

Let $e_0^* = \frac{1}{n\mu_y}(\Omega_y + \Omega_u)$ and $e_1^* = \frac{1}{n\mu_x}(\Omega_x + \Omega_v)$. In other words, $\hat{\bar{y}}^{*st} = (1 + e_0)\mu_y$ and $\bar{x}^{*st} = (1+e_1)\mu_x$, where $\hat{\bar{y}}^{*st} = \sum_{h=1}^{L}\pi_h(w_{1h}\bar{y}^*_{1h} + w_{2h}\hat{\bar{y}}^*_{2h})$ and $\bar{x}^{*st} = \sum_{h=1}^{L}\pi_h(w_{1h}\bar{x}^*_{1h} + w_{2h}\bar{x}^*_{2h})$ in the presence of measurement errors, and $\pi_h = N_h/N$.

Under the assumption of bivariate normality (Sukhatme et al.1970)[78]:

$$E(e_0^*) = E(e_1^*) = 0; \tag{V.5}$$

$$
\begin{aligned}
E(e_0^{*2}) =& \frac{1}{\mu_y^2} \sum_{h=1}^{L} \pi_h^2 [\theta_h(\sigma_{yh}^2 + \sigma_{uh}^2) + \lambda_h(\sigma_{y_{(2)h}}^2 + \sigma_{ph}^2) + \\
& \frac{W_{2h} f_h}{n_h} (\frac{\sigma_{sh}^2 W_h + \sigma_{Th}^2 W_h(\sigma_{y_{(2)h}}^2 + \sigma_{ph}^2 + \mu_{y_{(2)h}}^2)}{(\mu_{th} W_h + 1 + W_h)^2})],
\end{aligned} \tag{V.6}
$$

$$
E(e_1^{*2}) = \frac{1}{\mu_x^2} \sum_{h=1}^{L} \pi_h^2 [\theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma_{x_{(2)h}^2} + \sigma_{v_h}^2)], \tag{V.7}
$$

and

$$
E(e_0^* e_1^*) = \sum_{h=1}^{L} \pi_h^2 [\theta_h \rho_{yxh} \frac{\sigma_{yh}}{\mu_y} \frac{\sigma_{xh}}{\mu_x} + \lambda_h \rho_{zx(2)h} \frac{\sigma_{zh}}{\mu_z} \frac{\sigma_{x(2)h}}{\mu_x}], \tag{V.8}
$$

where $\theta_h = \frac{N_h - n_h}{N_h n_h}$, $\lambda_h = \frac{N_{2h}(f_h - 1)}{N_h n_h}$, $W_{2h} = \frac{N_{2h}}{N_h}$, and

$$
\rho_{zx_h} = \frac{\sigma_{yx_{(2)h}}}{\sqrt{1 + \frac{\sigma_{sh}^2 W_h + \sigma_{Th}^2 W_h(\sigma_{y_{(2)h}}^2 + \mu_{y_{(2)h}}^2)}{\sigma_{y_{(2)h}}^2}}}. \tag{V.9}
$$

The two existing mean estimators in the presence of measurement errors using the modified HH two-phase method under stratified random sampling are given by:

- The ordinary mean estimator is given by

$$
\hat{\mu}_{yw}^{st} = \sum_{h=1}^{L} \pi_h \hat{\bar{y}}_h^* = \sum_{h=1}^{L} \pi_h (w_{1h} \bar{y}_{1h}^* + w_{2h} \hat{\bar{y}}_{2h}^*), \tag{V.10}
$$

where $\pi_h = N_h/N$. It can be written as

$$
\hat{\mu}_{yw}^{st} = (1 + e_0^*) \mu_y. \tag{V.11}
$$

The difference between the ordinary mean estimator and the true mean can be written as

$$
\hat{\mu}_{yw}^{st} - \mu_y = e_0^* \mu_y. \tag{V.12}
$$

Taking square and then expected value on both side of (V.12), the MSE of $\hat{\mu}_{yw}^{st}$ is given by

$$MSE^*(\hat{\mu}_{yw}^{st}) = \sum_{h=1}^{L} \pi_h^2 [\theta_h(\sigma_{yh}^2 + \sigma_{uh}^2) + \lambda_h(\sigma_{y_{(2)h}}^2 + \sigma_{ph}^2) + G_h], \tag{V.13}$$

where $G_h = \frac{W_{2h} f_h}{n_h} [\frac{\sigma_{sh}^2 W_h + \sigma_{Th}^2 W_h (\sigma_{y_{(2)h}}^2 + \sigma_{ph}^2 + \mu_{y_{(2)h}}^2)}{(\mu_{th} W_h + 1 - W_h)^2}]$.

- A ratio estimator proposed by Gupta et al. (2014) is given by

$$\hat{\mu}_{rw}^{st} = \frac{\hat{\bar{y}}^{*st}}{\bar{x}^{*st}} \mu_x = \hat{R}_w^{st} \mu_x. \tag{V.14}$$

It can be written as

$$\begin{aligned} \hat{\mu}_{rw}^{st} &= \frac{(1+e_0^*)\mu_y}{(1+e_1^*)\mu_x} \mu_x \\ &= \mu_y (1+e_0^*)(1+e_1^*)^{-1} \\ &= \mu_y (1 - e_1^* + e_1^{*2} + e_0^* - e_0^* e_1^* + ...). \end{aligned} \tag{V.15}$$

Using second order approximation, the difference between the ratio estimator and the true mean can be written as

$$\hat{\mu}_{yw}^{st} - \mu_y = \mu_y(-e_1^* + e_1^{*2} + e_0^* - e_0^* e_1^*). \tag{V.16}$$

Taking square and then expected value on both side of (V.16), the MSE of $\hat{\mu}_{rw}^{st}$ is given by

$$\begin{aligned} MSE^*(\hat{\mu}_{rw}^{st}) = &\sum_{h=1}^{L} \pi_h^2 [\theta_h(\sigma_{yh}^2 + \sigma_{uh}^2 + R^2(\sigma_{xh}^2 + \sigma_{vh}^2) - 2R\rho_{yxh}\sigma_{yh}\sigma_{xh}) + \\ &\lambda_h(\sigma_{y_{(2)h}}^2 + \sigma_{ph}^2 + R^2(\sigma_{x_{(2)h}}^2 + \sigma_{v_h}^2) - 2R\rho_{zx_{(2)h}}\sigma_{zh}\sigma_{x_{(2)h}}) + G_h], \end{aligned} \tag{V.17}$$

where $R = \mu_y/\mu_x$.

The MSE of $\hat{\mu}_{yw}^{st}$ and $\hat{\mu}_{rw}^{st}$ without measurement error, may be obtained by putting $\sigma_{vh}^2 = \sigma_{uh}^2 = \sigma_{ph}^2 = 0$ in the equations (V.13)(V.17).

V.2. Generalized Estimator in the Presence of Measurement Errors and Non-response using ORRT Models under the Stratified Random Sampling Design

With this background, we use the generalized mean estimator used in Khalil et al. (2018) [34] and previous two Chapters. This estimator includes a wide variety of mean estimators as special cases. The non-response version under stratified random sampling is given by:

$$\hat{\mu}_{pw}^{st} = (\hat{\bar{y}}^{*st} + k(\mu_x - \bar{x}^{*st}))(\frac{\bar{D}^{st}}{\bar{d}^{st}})^v, \tag{V.18}$$

where $\hat{\bar{y}}^{*st} = \sum_{h=1}^{L} \pi_h(w_{1h}\bar{y}_{1h}^* + w_{2h}\hat{\bar{y}}_{2h}^*)$ is the ordinary mean estimator in (V.10), $\hat{\bar{x}}^{*st} = \sum_{h=1}^{L} \pi_h(w_{1h}\bar{x}_{1h}^* + w_{2h}\bar{x}_{2h}^*)$, $\bar{d} = \phi(\alpha^{st}\bar{x} + \beta^{st}) + (1-\phi)(\alpha^{st}\mu_x + \beta^{st})$, $\bar{D} = \alpha^{st}\mu_x + \beta^{st}$, k and v are suitable constants. ϕ is assumed to be an unknown constant whose value is to be determined from optimally considerations. α^{st} ($\alpha^{st} \neq 0$) and β^{st} are assumed to be some known parameters of the auxiliary variable X, such as coefficient of variation (C_x), kurtosis, and correlation coefficient (ρ_{yx}) etc. Please note here with different values of α^{st} and β^{st}, we can obtain various estimators. Also, with v=1 we get various ratio estimators and with v=-1 we get various product estimators.

V.2.1 Bias and MSE of the Generalized Mean Estimator

The generalized mean estimator will be studied under both modified HH two-phase sampling and measurement errors using stratified random sampling. According to the notations in Section V.1, it can be written as

$$\begin{aligned}\hat{\mu}_{pw}^{st} =&((1+e_0^*)\mu_y + k(\mu_x - (1+e_1^*)\mu_x))\\ &(\frac{\alpha^{st}\mu_x + \beta^{st}}{\phi(\alpha^{st}(1+e_1^*)\mu_x + \beta^{st}) + (1-\phi)(\alpha^{st}\mu_x + \beta^{st})})^v.\end{aligned} \tag{V.19}$$

Using Taylor's approximation and retaining terms of order up to 2, the difference between the generalized mean estimator and the true mean can be written as

$$\begin{aligned}&\hat{\mu}_{pw}^{st} - \mu_y\\ &= ((1+e_0^*)\mu_y + k(\mu_x - (1+e_1^*)\mu_x))(\frac{\alpha^{st}\mu_x + \beta^{st}}{\phi(\alpha^{st}(1+e_1^*)\mu_x + \beta^{st}) + (1-\phi)(\alpha^{st}\mu_x + \beta^{st})})^v\\ &\quad - \mu_y\\ &= e_0^*\mu_y + e_1^*[-k\mu_x + \mu_y v(\frac{-\alpha^{st}\phi\mu_x}{\alpha^{st}\mu_x + \beta^{st}})] + \frac{1}{2!}[2e_0^* e_1^* \mu_y v(\frac{-\alpha^{st}\phi\mu_x}{\alpha^{st}\mu_x + \beta^{st}}) +\\ &\quad e_1^{*2}(2kv\mu_x(\frac{\alpha^{st}\phi\mu_x}{\alpha^{st}\mu_x + \beta^{st}}) + v(v+1)\mu_y(\frac{-\alpha^{st}\phi\mu_x}{\alpha^{st}\mu_x + \beta^{st}})^2)].\end{aligned} \tag{V.20}$$

Taking expectation on both side of (V.20), the bias of the generalized mean estimator $\hat{\mu}_{pw}^{st}$, correct to second order of approximation, is given by

$$\begin{aligned}Bias(\hat{\mu}_{pw}^{st}) =& \sum_{h=1}^{L} \pi_h^2 \{\theta_h[(kH^{st} + \frac{v+1}{v}\mu_y H^{st2})(\sigma_{xh}^2 + \sigma_{vh}^2) - H^{st}\rho_{yxh}\sigma_{yh}\sigma_{xh}] +\\ &\lambda_h[(kH^{st} + \frac{v+1}{v}\mu_y H^{st2})(\sigma_{x(2)_h}^2 + \sigma_{v_h}^2) - H^{st}\mu_y\rho_{zx(2)h}\frac{\sigma_{zh}}{\mu_z}\sigma_{x(2)h}]\},\end{aligned} \tag{V.21}$$

where $H^{st} = \frac{\alpha^{st}\phi v}{\alpha^{st}\mu_x + \beta^{st}}$.

The bias of $\hat{\mu}_{pw}$ without measurement error may be obtained by setting $\sigma_v^2 = 0$ in equation (V.21).

To determine the expression for MSE of the generalized mean estimator (V.18), we take square of (V.20) on both sides and retain terms of order up to 2 which is given by

$$
\begin{aligned}
(\hat{\mu}_{pw}^{st} - \mu_y)^2 = e_0^{*2}\mu_y^2 + k^2\mu_x^2 e_1^{*2} + (H^{st}\mu_x\mu_y e_1^*)^2 - 2e_0^* e_1^* k\mu_x\mu_y - \\
2e_0^* e_1^* H^{st}\mu_x\mu_y^2 + 2e_1^{*2} k H^{st}\mu_x^2\mu_y.
\end{aligned}
\tag{V.22}
$$

Taking the expected value on both side of (V.22), the expression for MSE of $\hat{\mu}_{pw}^{st}$, correct to the first order approximation is given by

$$
\begin{aligned}
MSE^*(\hat{\mu}_{pw}^{st}) &= E(\hat{\mu}_{pw}^{st} - \mu_y)^2 \\
&\cong \sum_{h=1}^{L} \pi_h^2 [\theta_h(\sigma_{yh}^2 + k^2\sigma_{xh}^2 + \phi^2 v^2 R_{pw}^{st2}\sigma_{xh}^2 + 2k\phi v R_{pw}^{st}\sigma_{xh}^2 - 2k\rho_{yxh}\sigma_{xh}\sigma_{yh} \\
&\quad - 2\phi v R_{pw}^{st}\rho_{yxh}\sigma_{xh}\sigma_{yh}) + \lambda_h(\sigma_{y_{(2)h}}^2 + k^2\sigma_{x_{(2)h}}^2 + \phi^2 v^2 R_{pw}^{st2}\sigma_{x_{(2)h}}^2 \\
&\quad + 2k\phi v R_{pw}^{st}\sigma_{x_{(2)h}}^2 - 2k\rho_{zx_{(2)h}}\sigma_{zh}\sigma_{x_{(2)h}} - 2\phi v R_{pw}^{st}\mu_y\rho_{zx_{(2)h}}\sigma_{zh}\sigma_{x_{(2)h}}) \\
&\quad + \theta_h(\sigma_{uh}^2 + k^2\sigma_{vh}^2 + \phi^2 v^2 R_{pw}^{st2}\sigma_{vh}^2 + 2k\phi v R_{pw}^{st}\sigma_{vh}^2) + \\
&\quad \lambda_h(\sigma_{ph}^2 + k^2\sigma_{v_h}^2 + \phi^2 v^2 R_{pw}^{st2}\sigma_{v_h}^2 + 2k\phi v R_{pw}^{st}\sigma_{v_h}^2) + G_h] \\
&= \sum_{h=1}^{L} \pi_h^2 [\theta_h(\sigma_{yh}^2 + (k + \phi v R_{pw}^{st})^2\sigma_{xh}^2 - 2(k + \phi v R_{pw}^{st})\rho_{yx_h}\sigma_{xh}\sigma_{yh}) + \\
&\quad \lambda_h(\sigma_{y_{(2)h}}^2 + (k + \phi v R_{pw}^{st})^2\sigma_{x_{(2)h}}^2 - 2(k + \phi v R_{pw}^{st})\rho_{zx_{(2)h}}\sigma_{x_{(2)h}}\sigma_{zh}) + \\
&\quad \theta_h(\sigma_{uh}^2 + (k + \phi v R_{pw}^{st})^2\sigma_{vh}^2) + \lambda_h(\sigma_{ph}^2 + (k + \phi v R_{pw}^{st})^2\sigma_{vh}^2) + G_h,
\end{aligned}
\tag{V.23}
$$

where $R_{pw}^{st} = \frac{\alpha^{st}\mu_y}{\alpha^{st}\mu_x + \beta^{st}}$.

Minimization of the above expression (V.23) with respect to ϕ yields its optimum value as:

$$\phi_{opt} \cong \frac{\sum_{h=1}^{L} \pi_h^2[\theta_h(\rho_{yxh}\sigma_{xh}\sigma_{yh} - k(\sigma_{xh}^2 + \sigma_{vh}^2)) + \lambda_h(\mu_y \rho_{zx_h}\sigma_{zh}\sigma_{x_{(2)h}} - k(\sigma_{x_{(2)h}}^2 + \sigma_{v_h}^2))]}{vR_{pw}^{st}\sum_{h=1}^{L} \pi_h^2[\theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma_{x_{(2)h}}^2 + \sigma_{v_h}^2)]} \tag{V.24}$$

Substitution of ϕ_{opt} in MSE($\hat{\mu}_{pw}^{st}$) yields the minimum value as:

$$\begin{aligned} MSE_{min}^*(\hat{\mu}_{pw}^{st}) \cong \sum_{h=1}^{L} & \pi_h^2[\theta_h(\sigma_{yh}^2 + P^{st2}\sigma_{xh}^2 - 2P^{st}\rho_{yxh}\sigma_{xh}\sigma_{yh}) + \\ & \lambda_h(\sigma_{y_{(2)h}}^2 + P^{st2}\sigma_{x_{(2)h}}^2 - 2P^{st}\rho_{zx_{(2)h}}\mu_y\sigma_{zh}\sigma_{x_{(2)h}}) + \\ & \theta_h(\sigma_{uh}^2 + P^{st2}\sigma_{vh}^2) + \lambda_h(\sigma_{ph}^2 + P^{st2}\sigma_{v_h}^2) + G_h], \end{aligned} \tag{V.25}$$

where $P^{st} = \sum_{h=1}^{L} \pi_h^2 \frac{\theta_h\rho_{yxh}\sigma_{xh}\sigma_{yh} + \lambda_h\rho_{zx_{(2)h}}\sigma_{zh}\sigma_{x_{(2)h}}}{\theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma_{x_{(2)h}}^2 + \sigma_{v_h}^2)}$.

The expression for the minimized MSE of generalized estimator without measurement errors may be obtained by putting $\sigma_{uh}^2 = \sigma_{vh}^2 = \sigma_{ph}^2 = 0$ in the above expression, which gives

$$\begin{aligned} MSE_{min}(\hat{\mu}_{pw}^{st}) \cong \sum_{h=1}^{L} & \pi_h^2[\theta_h(\sigma_{yh}^2 + P^{st2}\sigma_{xh}^2 - 2P^{st}\rho_{yxh}\sigma_{xh}\sigma_{yh}) + \\ & \lambda_h(\sigma_{y_{(2)h}}^2 + P^{st2}\sigma_{x_{(2)h}}^2 - 2P^{st}\rho_{zx_{(2)h}}\sigma_{zh}\sigma_{x_{(2)h}}) + G_h], \end{aligned} \tag{V.26}$$

where $G_h = \frac{W_{2h}f_h}{n_h}[\frac{\sigma_{sh}^2 W_h + \sigma_{Th}^2 W_h(\sigma_{y_{(2)h}}^2 + \mu_{y_{(2)h}}^2)}{(\mu_{Th}W_h + 1 - W_h)^2}]$.

V.2.2 Efficiency Comparisons

Comparing MSE expressions of $\hat{\mu}_{yw}^{st}$ (V.13), $\hat{\mu}_{rw}^{st}$ (V.17), and $\hat{\mu}_{pw}^{st}$ (V.25) with measurement errors, it can be verified that

- $MSE^*_{min}(\hat{\mu}^{st}_{pw}) < MSE^*(\hat{\mu}^{st}_{yw})$ if

$$\begin{aligned} & MSE^*_{min}(\hat{\mu}^{st}_{pw}) - MSE^*(\hat{\mu}^{st}_{yw}) \\ &= -\sum_{h=1}^{L} \pi_h^2 \frac{(\theta_h \rho_{yxh}\sigma_{xh}\sigma_{yh} + \lambda_h \rho_{zx_{(2)h}}\sigma_{zh}\sigma_{x_{(2)h}})^2}{\theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma^2_{x_{(2)h}} + \sigma^2_{vh})} < 0, \end{aligned} \tag{V.27}$$

- $MSE^*_{min}(\hat{\mu}^{st}_{pw}) < MSE^*(\hat{\mu}^{st}_{rw})$ if

$$\begin{aligned} & MSE^*_{min}(\hat{\mu}^{st}_{pw}) - MSE^*(\hat{\mu}^{st}_{rw}) \\ =& \sum_{h=1}^{L} \pi_h^2 [(P^{st2} - R^2)(\theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma^2_{x_{(2)h}} + \sigma^2_{v_h})) \\ & - 2(P^{st} + R)(\theta_h \rho_{yx_h}\sigma_{xh}\sigma_{yh} + \lambda \rho_{zx_{(2)h}}\sigma_{zh}\sigma^2_{x_{(2)h}})] < 0; \end{aligned} \tag{V.28}$$

In other words, if

$$\begin{aligned} & \sum_{h=1}^{L} \pi_h^2 \frac{(P^{st2} - R^2)(\theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma^2_{x_{(2)h}} + \sigma^2_{v_h}))}{2(P^{st} + R)(\theta_h \rho_{yx_h}\sigma_{xh}\sigma_{yh} + \lambda \rho_{zx_{(2)h}}\sigma_{zh}\sigma^2_{x_{(2)h}})} \\ =& \sum_{h=1}^{L} \pi_h^2 \frac{(P^{st} - R)(\theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda(\sigma^2_{x_{(2)h}} + \sigma^2_{vh}))}{2(\theta_h \rho_{yxh}\sigma_{xh}\sigma_{yh} + \lambda_h \rho_{zx_{(2)h}}\sigma_{zh}\sigma_{x_{(2)h}})} \\ =& \frac{P^{st} - R}{2P^{st}} \\ =& \frac{1}{2} - \frac{R}{2P^{st}} \\ =& \frac{1}{2} - \frac{\mu_y}{2\mu_x} \sum_{h=1}^{L} \pi_h^2 \frac{\theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma^2_{x_{(2)h}} + \sigma^2_{v_h})}{\theta_h \rho_{yx_h}\sigma_{xh}\sigma_{yh} + \lambda_h \rho_{zx_{(2)h}}\sigma_{zh}\sigma_{x_{(2)h}}} < 1, \end{aligned} \tag{V.29}$$

and

- $MSE^*_{min}(\hat{\mu}^{st}_{rw}) < MSE^*(\hat{\mu}^{st}_{yw})$ if

$$\begin{aligned} MSE^*(\hat{\mu}^{st}_{rw}) - MSE^*(\hat{\mu}^{st}_{yw}) = &\sum_{h=1}^{L} \pi_h^2 [R^2(\theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma^2_{x_{(2)h}} + \sigma^2_{v_h})) \\ &- 2R(\theta_h \rho_{yx_h} \sigma_{xh} \sigma_{yh} + \lambda_h \rho_{zx_{(2)h}} \sigma_{zh} \sigma_{x_{(2)h}})] < 0; \end{aligned} \tag{V.30}$$

In other words, if

$$\begin{aligned} &\sum_{h=1}^{L} \pi_h^2 \frac{R^2\theta_h(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma^2_{x_{(2)h}} + \sigma^2_{v_h})}{2R\theta_h \rho_{yx_h} \sigma_{xh} \sigma_{yh} + \lambda_h \rho_{zx_{(2)h}} \sigma_{zh} \sigma_{x_{(2)h}}} \\ =&\frac{\mu_y}{2\mu_x} \sum_{h=1}^{L} \pi_h^2 \frac{\theta(\sigma_{xh}^2 + \sigma_{vh}^2) + \lambda_h(\sigma^2_{x_{(2)h}} + \sigma^2_{vh})}{\theta \rho_{yxh} \sigma_{xh} \sigma_{yh} + \lambda_h \rho_{zx_{(2)h}} \sigma_{zh} \sigma_{x_{(2)h}}} < 1. \end{aligned} \tag{V.31}$$

The conditions (V.27) and (V.29) always hold true. From (V.31), the ratio estimator is more efficient than the ordinary mean estimator only if the measurement error on auxiliary variable X (σ^2_{vh}) is small and X and Y are highly correlated in each stratum.

V.3. Simulation Study

We will evaluate the performance of the generalized mean estimator under non-response and measurement errors using stratified random sampling with the other two estimators by a simulation study in this section. In the generalized mean estimator, we choose v and k to be 1, $\alpha = 1$, $\beta = 0$, and ϕ to be its optimum value. The scrambling variable S is taken to be a normal variate with mean equal to zero and vary variance ($0.2{*}\sigma_x^2$, $0.5{*}\sigma_x^2$, σ_x^2). And T is also taken to be a normal variate but with mean equal to one and varying variances (0, 0.5, 1). The measurement errors of X have a normal distribution with mean zero in both phases; the measurement errors of Y in the first phase and Z in the second phase have a normal distribution with mean zero. We use different variances for measurement errors.

We consider three bivariate normal distributions with different covariance matrices to represent the distribution of Y and X in three strata. We assume each stratum with size of 2000 and take a sample of size 200 using SRSWOR from each stratum. We assume the first phase response rate in each stratum is 40%. This means in the first phase only 80 (n_1) subjects provide a response to the survey question and 120 (n_2) of them do not. In the second phase, we take another sample ($n_s = \frac{n_2}{f}$) from non-respondent group by using f=2, 3, 4, respectively. Different response rates of 20%, 40% and 60% also compared in the simulation study. The three strata have covariance matrices Σ as given below:

$$Stratum1 \quad \mu = \begin{bmatrix} 10 \\ 6 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 16 & 7.8384 \\ 7.8384 & 6 \end{bmatrix}, \quad \rho_{yx} = 0.8$$

$$Stratum2 \quad \mu = \begin{bmatrix} 9 \\ 5 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 16 & 6.2610 \\ 6.2610 & 5 \end{bmatrix}, \quad \rho_{yx} = 0.7$$

$$Stratum3 \quad \mu = \begin{bmatrix} 7 \\ 4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 16 & 6.8586 \\ 6.8586 & 6 \end{bmatrix}, \quad \rho_{yx} = 0.7$$

The real parameters of the set of 5000 data points we generated using R are very close to the parameter values in (A) but not exactly same. For the simulation study, we used parameter values in (B) and not those in (A).

Stratum 1

$$\mu_x = 6,\ \sigma_x^2 = 6,\ \mu_y = 10,\ \sigma_y^2 = 16,\ \rho_{yx} = 0.8 \tag{A}$$

$$\mu_x = 6.0512,\ \sigma_x^2 = 5.9809,\ \mu_y = 9.9802,\ \sigma_y^2 = 16.0583,\ \rho_{yx} = 0.8129 \tag{B}$$

Stratum 2

$$\mu_x = 5,\ \sigma_x^2 = 5,\ \mu_y = 9,\ \sigma_y^2 = 16,\ \rho_{yx} = 0.7 \tag{A}$$

$$\mu_x = 5.0968,\ \sigma_x^2 = 5.0155,\ \mu_y = 9.2189,\ \sigma_y^2 = 16.2913,\ \rho_{yx} = 0.6887 \tag{B}$$

Stratum 3

$$\mu_x = 4,\ \sigma_x^2 = 6,\ \mu_y = 7,\ \sigma_y^2 = 16,\ \rho_{yx} = 0.7 \tag{A}$$

$$\mu_x = 3.9674,\ \sigma_x^2 = 5.9354,\ \mu_y = 6.8833,\ \sigma_y^2 = 15.5977,\ \rho_{yx} = 0.7057 \tag{B}$$

Coding for the simulations was done in R and results are averaged over 5,000 iterations. The empirical MSE of the estimator $\hat{\mu}_y$ is computed by

$$MSE^*(\hat{\mu}_w) = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\mu}_w - \mu)^2, \tag{V.32}$$

where $\hat{\mu}_w = \hat{\mu}_{yw}^{st},\ \hat{\mu}_{rw}^{st},\ \hat{\mu}_{pw}^{st}$. Here, μ is the population mean of the sensitive study variable. The percent relative efficiencies (PREs) of the estimators $(\hat{\mu}_w)$ with respect to the ordinary mean estimator $(\hat{\mu}_{yw}^{st}\ )$ is defined as

$$PRE = \frac{MSE^*(\hat{\mu}_{yw}^{st})}{MSE^*(\hat{\mu}_w)} * 100. \tag{V.33}$$

We will also use the unified measure δ of efficiency and privacy as defined in Gupta et al. (2018)[24]. It is given by

$$\delta = \frac{MSE^*(\hat{\mu}_w)}{\Delta_{DP}}. \tag{V.34}$$

In (V.34), MSE is used in place of Var(.) to account for biased estimators. And Δ_{DP} is calculated by $\Delta_{DP} = \sum_{h=1}^{L}\pi_h\Delta_{DPh}$, where Δ_{DPh} is the privacy level in the h^{th} stratum.

Table V.1. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators under Stratified Random Sampling when Response Rate = 40% , σ_v^2 = σ_u^2= σ_p^2 = 1, f = 2 and σ_s^2 = 0.2*σ_x^2.

Est.	W	σ_T^2	MSE		PRE		δ
			Without ME	With ME	Without ME	With ME	
$\hat{\mu}_{yw}^{st}$	0.5	0	**0.0408**	**0.0433**	**100.0000**	**100.0000**	**0.0371**
			0.0412	0.0437	100.0000	100.0000	0.0374
		0.5	**0.0864**	**0.0890**	**100.0000**	**100.0000**	**0.0015**
			0.0910	0.0913	100.0000	100.0000	0.0016
		1	**0.1319**	**0.1344**	**100.0000**	**100.0000**	**0.0012**
			0.1259	0.1335	100.0000	100.0000	0.0012
	0.8	0	**0.0415**	**0.0440**	**100.0000**	**100.0000**	**0.0377**
			0.0421	0.0438	100.0000	100.0000	0.0375
		0.5	**0.1142**	**0.1167**	**100.0000**	**100.0000**	**0.0020**
			0.1191	0.1169	100.0000	100.0000	0.0020
		1	**0.1869**	**0.1894**	**100.0000**	**100.0000**	**0.0017**
			0.1859	0.1835	100.0000	100.0000	0.0016
	1	0	**0.0420**	**0.0445**	**100.0000**	**100.0000**	**0.0381**
			0.0422	0.0432	100.0000	100.0000	0.0370
		0.5	**0.1325**	**0.1350**	**100.0000**	**100.0000**	**0.0023**
			0.1361	0.1337	100.0000	100.0000	0.0023
		1	**0.2233**	**0.2258**	**100.0000**	**100.0000**	**0.0020**
			0.2261	0.2230	100.0000	100.0000	0.0019
$\hat{\mu}_{rw}^{st}$	0.5	0	**0.0264**	**0.0364**	**154.5455**	**118.9560**	**0.0312**
			0.0268	0.0365	153.7313	119.7260	0.0312
		0.5	**0.0711**	**0.0810**	**121.5190**	**109.8765**	**0.0014**
			0.0693	0.0820	131.3131	111.3415	0.0014
		1	**0.1161**	**0.1260**	**113.6090**	**106.6667**	**0.0011**
			0.1193	0.1219	105.5323	109.5160	0.0011
	0.8	0	**0.0271**	**0.0370**	**153.1365**	**118.9189**	**0.0317**
			0.0277	0.0357	151.9856	122.6891	0.0306
		0.5	**0.0985**	**0.1084**	**115.9391**	**107.6568**	**0.0019**

			0.1022	0.1103	116.5362	105.9837	0.0019
		1	**0.1706**	**0.1806**	**109.5545**	**104.8726**	**0.0016**
			0.1691	0.1723	109.9349	106.5003	0.0015
	1	0	**0.0275**	**0.0375**	**152.7273**	**118.6667**	**0.0321**
			0.0283	0.0369	149.1166	117.0732	0.0316
		0.5	**0.1166**	**0.1266**	**113.6364**	**106.6351**	**0.0022**
			0.1219	0.1303	111.6489	102.6094	0.0023
		1	**0.2069**	**0.2168**	**107.9265**	**104.1513**	**0.0019**
			0.1999	0.2122	113.1066	105.0895	0.0019
$\hat{\mu}_{pw}^{st}$	0.5	0	**0.0214**	**0.0269**	**190.6542**	**160.9665**	**0.0230**
			0.0202	0.0259	203.9604	168.7259	0.0222
		0.5	**0.0665**	**0.0720**	**129.9248**	**123.6111**	**0.0012**
			0.0668	0.0730	136.2275	125.0685	0.0013
		1	**0.1117**	**0.1173**	**118.0842**	**114.5780**	**0.0010**
			0.1092	0.1147	115.2930	116.3906	0.0010
	0.8	0	**0.0221**	**0.0275**	**187.7828**	**160.0000**	**0.0235**
			0.0224	0.0275	187.9464	159.2727	0.0235
		0.5	**0.0940**	**0.0996**	**121.4894**	**117.1687**	**0.0017**
			0.0982	0.1009	121.2831	115.8573	0.0017
		1	**0.1664**	**0.1720**	**112.3197**	**110.1163**	**0.0015**
			0.1633	0.1651	113.8396	111.1448	0.0014
	1	0	**0.0225**	**0.0280**	**186.6667**	**158.9286**	**0.0240**
			0.0229	0.0276	184.2795	156.5217	0.0236
		0.5	**0.1123**	**0.1178**	**117.9875**	**114.6010**	**0.0020**
			0.1174	0.1148	115.9284	116.4634	0.0020
		1	**0.1927**	**0.2083**	**115.8796**	**108.4013**	**0.0018**
			0.1929	0.2043	117.2110	109.1532	0.0018

Table V.2. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators under Stratified Random Sampling when Response Rate = 40% , σ_v^2 = σ_u^2= σ_p^2 = 1, f = 2 and σ_s^2 = 0.5*σ_x^2.

Est.	W	σ_T^2	MSE		PRE		δ
			Without ME	With ME	Without ME	With ME	
$\hat{\mu}_{yw}^{st}$	0.5	0	**0.0425**	**0.0450**	**100.0000**	**100.0000**	**0.0158**
			0.0429	0.0456	100.0000	100.0000	0.0160
		0.5	**0.0882**	**0.0907**	**100.0000**	**100.0000**	**0.0015**
			0.0929	0.0932	100.0000	100.0000	0.0016
		1	**0.1337**	**0.1362**	**100.0000**	**100.0000**	**0.0012**
			0.1279	0.1357	100.0000	100.0000	0.0011
	0.8	0	**0.0443**	**0.0468**	**100.0000**	**100.0000**	**0.0165**
			0.0452	0.0465	100.0000	100.0000	0.0164
		0.5	**0.1169**	**0.1294**	**100.0000**	**100.0000**	**0.0022**
			0.1224	0.1301	100.0000	100.0000	0.0022
		1	**0.1896**	**0.1921**	**100.0000**	**100.0000**	**0.0016**
			0.1890	0.1868	100.0000	100.0000	0.0016
	1	0	**0.0454**	**0.0479**	**100.0000**	**100.0000**	**0.0168**
			0.0458	0.0469	100.0000	100.0000	0.0165
		0.5	**0.1359**	**0.1384**	**100.0000**	**100.0000**	**0.0023**
			0.1404	0.1378	100.0000	100.0000	0.0023
		1	**0.2268**	**0.2293**	**100.0000**	**100.0000**	**0.0019**
			0.2201	0.2269	100.0000	100.0000	0.0019
$\hat{\mu}_{rw}^{st}$	0.5	0	**0.0281**	**0.0380**	**151.2456**	**118.4211**	**0.0134**
			0.0255	0.0382	168.2353	119.3717	0.0134
		0.5	**0.0728**	**0.0827**	**121.1538**	**109.6735**	**0.0014**
			0.0710	0.0841	130.8451	110.8205	0.0014
		1	**0.1178**	**0.1277**	**113.4975**	**106.6562**	**0.0011**
			0.1112	0.1237	115.0180	109.7009	0.0010
	0.8	0	**0.0297**	**0.0397**	**149.1582**	**117.8841**	**0.0140**
			0.0295	0.0372	153.2203	125.0000	0.0131
		0.5	**0.1012**	**0.1111**	**115.5138**	**116.4716**	**0.0019**

			0.1056	0.1130	115.9091	115.1327	0.0019
		1	**0.1733**	**0.1833**	**109.4057**	**104.8009**	**0.0015**
			0.1691	0.1753	111.7682	106.5602	0.0015
	1	0	**0.0308**	**0.0408**	**147.4026**	**117.4020**	**0.0144**
			0.0307	0.0393	149.1857	119.3384	0.0138
		0.5	**0.1200**	**0.1299**	**113.2500**	**106.5435**	**0.0022**
			0.1259	0.1360	111.5171	101.3235	0.0023
		1	**0.2103**	**0.2203**	**107.8459**	**104.0853**	**0.0019**
			0.1975	0.2161	111.4430	104.9977	0.0018
$\hat{\mu}_{pw}^{st}$	0.5	0	**0.0231**	**0.0286**	**183.9827**	**157.3427**	**0.0101**
			0.0218	0.0267	196.7890	170.7865	0.0094
		0.5	**0.0682**	**0.0737**	**129.3255**	**123.0665**	**0.0012**
			0.0686	0.0749	135.4227	124.4326	0.0013
		1	**0.1134**	**0.1190**	**117.9012**	**114.4538**	**0.0010**
			0.1071	0.1166	119.4211	116.3808	0.0010
	0.8	0	**0.0248**	**0.0302**	**178.6290**	**154.9669**	**0.0106**
			0.0233	0.0292	193.9914	159.2466	0.0103
		0.5	**0.0967**	**0.1023**	**120.8893**	**126.4907**	**0.0017**
			0.1005	0.1028	121.7910	126.5564	0.0017
		1	**0.1691**	**0.1747**	**112.1230**	**109.9599**	**0.0015**
			0.1663	0.1682	113.6500	111.0583	0.0014
	1	0	**0.0259**	**0.0314**	**175.2896**	**152.5478**	**0.0110**
			0.0243	0.0301	188.4774	155.8140	0.0106
		0.5	**0.1156**	**0.1212**	**117.5606**	**114.1914**	**0.0020**
			0.1215	0.1216	115.5556	113.3224	0.0021
		1	**0.1962**	**0.2118**	**115.5963**	**108.2625**	**0.0018**
			0.1956	0.2082	112.5256	108.9817	0.0018

Table V.3. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators under Stratified Random Sampling when Response Rate = 40% , σ_v^2 = σ_u^2= σ_p^2 = 1, f = 2 and σ_s^2 = 1*σ_x^2.

Est.	W	σ_T^2	MSE		PRE		δ
			Without ME	With ME	Without ME	With ME	
$\hat{\mu}_{yw}^{st}$	0.5	**0**	**0.0454**	**0.0479**	**100.0000**	**100.0000**	**0.0410**
			0.0459	0.0486	100.0000	100.0000	0.0416
		0.5	**0.0910**	**0.0935**	**100.0000**	**100.0000**	**0.0016**
			0.0939	0.0962	100.0000	100.0000	0.0017
		1	**0.1365**	**0.1390**	**100.0000**	**100.0000**	**0.0012**
			0.1369	0.1390	100.0000	100.0000	0.0012
	0.8	**0**	**0.0489**	**0.0514**	**100.0000**	**100.0000**	**0.0440**
			0.0502	0.0511	100.0000	100.0000	0.0437
		0.5	**0.1215**	**0.1240**	**100.0000**	**100.0000**	**0.0021**
			0.1255	0.1250	100.0000	100.0000	0.0022
		1	**0.1942**	**0.1967**	**100.0000**	**100.0000**	**0.0017**
			0.1940	0.1921	100.0000	100.0000	0.0017
	1	**0**	**0.0512**	**0.0537**	**100.0000**	**100.0000**	**0.0460**
			0.0516	0.0528	100.0000	100.0000	0.0452
		0.5	**0.1416**	**0.1441**	**100.0000**	**100.0000**	**0.0025**
			0.1469	0.1441	100.0000	100.0000	0.0025
		1	**0.2325**	**0.2350**	**100.0000**	**100.0000**	**0.0021**
			0.2265	0.2330	100.0000	100.0000	0.0020
$\hat{\mu}_{rw}^{st}$	0.5	0	**0.0308**	**0.0408**	**147.4026**	**117.4020**	**0.0349**
			0.0292	0.0391	157.1918	124.2967	0.0335
		0.5	**0.0756**	**0.0855**	**120.3704**	**109.3567**	**0.0015**
			0.0737	0.0873	127.4084	110.1947	0.0015
		1	**0.1206**	**0.1306**	**113.1841**	**106.4319**	**0.0011**
			0.1193	0.1286	114.7527	108.0871	0.0011
	0.8	0	**0.0341**	**0.0441**	**143.4018**	**116.5533**	**0.0378**
			0.0341	0.0435	147.2141	117.4713	0.0372
		0.5	**0.1056**	**0.1156**	**115.0568**	**107.2664**	**0.0020**

			0.1099	0.1176	114.1947	106.2925	0.0020
		1	**0.1779**	**0.1878**	**109.1625**	**104.7391**	**0.0016**
			0.1739	0.1802	111.5584	106.6038	0.0016
		0	**0.0364**	**0.0463**	**140.6593**	**115.9827**	**0.0396**
			0.0354	0.0449	145.7627	117.5947	0.0384
	1	0.5	**0.1256**	**0.1356**	**112.7389**	**106.2684**	**0.0023**
			0.1222	0.1420	120.2128	101.4789	0.0025
		1	**0.2160**	**0.2260**	**107.6389**	**103.9823**	**0.0020**
			0.2133	0.2222	106.1885	104.8605	0.0019
		0	**0.0259**	**0.0314**	**175.2896**	**152.5478**	**0.0269**
			0.0246	0.0306	186.5854	158.8235	0.0262
	0.5	0.5	**0.0710**	**0.0765**	**128.1690**	**122.2222**	**0.0013**
			0.0714	0.0780	131.5126	123.3333	0.0013
		1	**0.1163**	**0.1218**	**117.3689**	**114.1215**	**0.0011**
			0.1162	0.1198	117.8141	116.0267	0.0010
		0	**0.0293**	**0.0348**	**166.8942**	**147.7011**	**0.0298**
			0.0280	0.0336	179.2857	152.0833	0.0288
$\hat{\mu}_{pw}^{st}$	0.8	0.5	**0.1012**	**0.1068**	**120.0593**	**116.1049**	**0.0018**
			0.1017	0.1025	123.4022	121.9512	0.0018
		1	**0.1737**	**0.1793**	**111.8020**	**109.7044**	**0.0016**
			0.1712	0.1733	113.3178	110.8482	0.0015
		0	**0.0315**	**0.0370**	**162.5397**	**145.1351**	**0.0317**
			0.0301	0.0358	171.4286	147.4860	0.0306
	1	0.5	**0.1213**	**0.1269**	**116.7354**	**113.5540**	**0.0022**
			0.1278	0.1246	114.9452	115.6501	0.0022
		1	**0.2119**	**0.2175**	**109.7216**	**108.0460**	**0.0019**
			0.2116	0.2143	107.0416	108.7261	0.0019

Tables V.1, V.2, V.3 present the theoretical and empirical MSEs and PREs of the ORRT mean estimators under stratified random sampling when all the variances of measurement errors (σ_v^2, σ_u^2 and σ_p^2) are set equal to 1 and response rate in Phase I is set equal to 40% with different variances of S ($0.2^*\sigma_x^2$, $0.5^*\sigma_x^2$, $1^*\sigma_x^2$) , respectively. Also, the mean estimation is less efficient as the variance of S increases in the presence of non-response. This is consistent with the theoretical results. As mentioned earlier, larger variance of S introduces more penalty for using RRT models.

For all three tables, the MSE of the mean estimators increases as W increases under non-response and measurement errors. For example, in Table V.3, the MSE of the generalized mean estimator increased from 0.1218 to 0.2175 as the sensitivity level increased from 0.5 to 1 when variance of T is equal to 1. It indicates that the ORRT model is more efficient when some of the respondents feel the survey question is not sensitive. In addition, as the variance of T increases, the MSE increases while δ decreases with a reasonably small value of σ_T^2.

Table V.4. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators under the Conditions of $\sigma_v^2 = \sigma_u^2 = \sigma_p^2 = 1$, 5, 10 and Stratified Random Sampling when Response Rate = 40%, W = 0.8, and $\sigma_T^2 = 0.5^*\sigma_x^2$.

Est.	f	MSE			PRE		
		1	5	10	1	5	10
$\hat{\mu}_{yw}^{st}$	2	**0.1294**	**0.1395**	**0.1417**	**100.0000**	**100.0000**	**100.0000**
		0.1301	0.1445	0.1443	100.0000	100.0000	100.0000
	3	**0.1747**	**0.1888**	**0.2059**	**100.0000**	**100.0000**	**100.0000**
		0.1770	0.1956	0.2053	100.0000	100.0000	100.0000
	4	**0.2300**	**0.2481**	**0.2701**	**100.0000**	**100.0000**	**100.0000**
		0.2363	0.2482	0.2765	100.0000	100.0000	100.0000
$\hat{\mu}_{rw}^{st}$	2	**0.1111**	**0.1509**	**0.1868**	**116.4716**	**92.4453**	**75.8565**
		0.1130	0.1524	0.1898	115.1327	94.8163	76.0274
	3	**0.1645**	**0.2202**	**0.2838**	**106.2006**	**85.7402**	**72.5511**
		0.1692	0.2207	0.2831	104.6099	88.6271	72.5185

	4	**0.2180**	**0.2895**	**0.3707**	**105.5046**	**85.6995**	**72.8622**
		0.2182	0.2855	0.3679	108.2951	86.9352	75.1563
$\hat{\mu}_{pw}^{st}$	2	**0.1023**	**0.1288**	**0.1342**	**126.4907**	**108.3075**	**105.5887**
		0.1028	0.1282	0.1323	126.5564	112.7145	109.0703
	3	**0.1515**	**0.1743**	**0.1957**	**115.3135**	**108.3190**	**105.2121**
		0.1528	0.1770	0.1991	115.8377	110.5085	103.1140
	4	**0.2007**	**0.2298**	**0.2572**	**114.5989**	**107.9634**	**105.0156**
		0.1986	0.2329	0.2575	118.9829	106.5693	107.3786

Table V.5. Theoretical (**bold**) and Empirical MSEs/PREs of the ORRT Estimators under the Conditions of Response Rate = 20%, 40%, 60% and Stratified Random Sampling when $\sigma_v^2 = \sigma_u^2 = \sigma_p^2 = 1$, W = 0.8, and $\sigma_T^2 = 0.5 * \sigma_x^2$.

Response Rate	f	MSE			PRE		
		20%	40%	60%	20%	40%	60%
$\hat{\mu}_{yw}^{st}$	2	**0.1541**	**0.1294**	**0.0882**	**100.0000**	**100.0000**	**100.0000**
		0.1576	0.1301	0.0864	100.0000	100.0000	100.0000
	3	**0.2297**	**0.1747**	**0.1252**	**100.0000**	**100.0000**	**100.0000**
		0.2292	0.1770	0.1259	100.0000	100.0000	100.0000
	4	**0.3053**	**0.2300**	**0.1621**	**100.0000**	**100.0000**	**100.0000**
		0.3035	0.2363	0.1624	100.0000	100.0000	100.0000
$\hat{\mu}_{rw}^{st}$	2	**0.1447**	**0.1111**	**0.0806**	**106.4962**	**116.4716**	**109.4293**
		0.1448	0.1130	0.0822	108.8398	115.1327	105.1095
	3	**0.2174**	**0.1645**	**0.1164**	**105.6578**	**106.2006**	**107.5601**
		0.2119	0.1692	0.1209	108.1642	104.6099	104.1356
	4	**0.2901**	**0.2180**	**0.1522**	**105.2396**	**105.5046**	**106.5046**
		0.2937	0.2182	0.1548	103.3367	108.2951	104.9096
$\hat{\mu}_{pw}^{st}$	2	**0.1346**	**0.1023**	**0.0732**	**114.4874**	**126.4907**	**120.4918**
		0.1308	0.1028	0.0778	120.4893	126.5564	111.0825
	3	**0.2017**	**0.1515**	**0.1062**	**113.8820**	**115.3135**	**117.8908**
		0.2040	0.1528	0.1128	112.3529	115.8377	111.6135
	4	**0.2689**	**0.2007**	**0.1391**	**113.5366**	**114.5989**	**116.5349**
		0.2681	0.1986	0.1455	113.2040	118.9829	111.6151

Table V.4 presents the theoretical and empirical MSEs and PREs of the ORRT mean estimators under stratified random sampling and different variances of measurement errors (1, 5, 10) when the sensitivity level W is equal to 0.8, variance of T is equal to 0.5 and response rate in Phase I is equal to 40%. As the measurement errors increase, the MSE of each mean estimator increases. For example, the MSE of the ratio estimator increased from 0.1645 to 0.2838 as the variance of measurement errors increased from 1 to 10 when the value of f is 3. It is obvious that larger measurement errors have larger negative impact on mean estimation under non-response using stratified random sampling, as was the case with simple random sampling.

Also, from Tables V.1, V.2, V.3 and V.4, it is clear that the generalized mean estimator $\hat{\mu}_{pw}$ is more efficient than the other two mean estimators even when very large measurement errors are present. However, the ratio estimator $\hat{\mu}_{rw}$ becomes less efficient than the ordinary mean estimator $\hat{\mu}_{yw}$ as the measurement errors increase. For example in Table V.4, the MSE of the generalized mean estimator 0.1288 is less than the MSE of the ordinary mean estimator 0.1395 when the variance of measurement errors is 5 and the value of f is 2. However, the MSE of the ratio estimator 0.1509 is larger than other two estimators. Similar to mean estimation under simple random sampling, measurement error on X makes the ratio estimator less efficient than the ordinary mean estimator unless the variance of measurement error on X is small because measurement errors exist only on Y for the ordinary mean estimator while in the ratio estimator measurement errors exist on both X and Y. At the same time, this result shows the superiority of the generalized mean estimator in the presence of measurement errors and non-response using stratified random sampling because it is not affected as poorly as the ratio estimator is by measurement errors on X.

Table V.5 presents the theoretical and empirical MSEs and PREs of the ORRT mean estimators under stratified random sampling and different response rates when the variance of measurement errors is equal to 1, sensitivity level W is equal to 0.8, and variance of T is equal to 0.5. The efficiency of each estimator gets better as the response rate increases. In other words, the larger the sample we collect from the first call in each stratum, the higher is the efficiency of the mean estimation.

In addition, from both Tables V.4 and V.5, the efficiency of each estimator gets better as the value of f decreases. For example, the MSE of the generalized mean estimator decreased from 0.2007 to 0.1023 as the value of f decreased from 4 to 2 when the variance of measurement errors is 1. It is reasonable because smaller f value means we obtain a larger sample from the second call in each stratum and the mean estimation is more efficient when a larger sample is used.

V.4. Concluding Chapter Remarks

The main contribution in this chapter is the re-examination of Chapter IV but under stratified random sampling. Under stratified random sampling, all the conclusions we made under simple random sampling still hold true. From the theoretical conditions (V.27) (V.29) (V.31) and simulation results, the generalized mean estimator is more efficient than the ordinary mean estimator and the ratio estimator when the measurement errors and non-response are present, while the ratio estimator is less efficient than the ordinary mean estimator when the measurement errors on X are large. The reason is the ordinary mean estimator is not affected by the measurement error in X. Even though the generalized mean estimator is also affected by measurement error on X, the use of the regression term was able to overcome the measurement error burden on X.

CHAPTER VI

CONCLUDING REMARKS AND FUTURE DIRECTIONS

Mean estimation of a sensitive variable under measurement errors and non-response using both simple random sampling and stratified random sampling is studied in this dissertation. The empirical results are in good agreement with the corresponding theoretical conclusions. The simple additive RRT is more efficient if we ignore privacy issue, but the general linear combination RRT model is better if we examine the performance of various estimators with respect to the unified measure of efficiency and privacy. The MSE of all mean estimators increases as W increases under all conditions, which shows that the ORRT model leads to better results than non-optional model. The generalized mean estimator always performs better than both the ordinary mean estimator and ratio estimator under measurement errors and non-response. The ratio estimator is more efficient than the ordinary mean estimator when the study variable and the auxiliary variable are highly correlated and the measurement errors on the auxiliary variable are small.

For future studies, one can consider mean estimation of a sensitive variable using different sampling methods, such as unequal probability sampling. Also, instead of estimating mean, one can estimate other parameters like the variance and the distribution function.

BIBLIOGRAPHY

[01] Ahmed, S., Shabbir, J., and Gupta, S.(2017) Use of scrambled response model in estimating the finite population mean in presence of non-reponse when coefficient of variation is known. *Communications in Statistics - Theory and Methods*, 46(17):8435-8449.

[02] Allen, J., Singh, H.P. and Smarandache, F. (2003). A family of estimators of population mean using multi-auxiliary information in presence of measurement errors. *International Journal of Social Economics information* , 30(7):837-849.

[03] Azeem, M. and Hanif, M. (2017). Joint influence of measurement error and non-response on estimation of population mean. *Communications in Statistics - Theory and Methods*, 46(4):1679-1693.

[04] Blattman, C., Gonwa, T., Jamison, J., Rodrigues, K., Sheridan, M. (2014). Measuring the measurement error: A method to qualitatively validate sensitive survey data. *ournal of Development Economics*, 120(C):99-112.

[05] Chen X. and Du Q. and Jin Z. et al. (2014). The randomized response technique application in the survey of homosexual commercial sex among men in Beijing. *Iranian Journal of Public Health*, 43(4):416-422.

[06] Chhabra, A. and Dass, B.K. and Gupta, S. (2016). Estimating prevalence of sexual abuse by an acquaintance with an optional unrelated question RRT model. *North Carolina Journal of Mathematics and Statistics*, 2:1-9

[07] Cochran. W.G. (1977). Sampling Techniques. 3rd Edition. John Wiley, New York.

[08] Diana, G., and Perri, P.F. (2011). A class of estimators for quantitative sensitive data. *Statistical Papers*, 52:633–650.

[09] Diana, G., Riaz, S., and Shabbir, J. (2014) Hansen and Hurwitz estimator with scrambled response on the second call. *J. Appl. Stat.*, 41(3):596–611.

[10] Eichhorn, B.H. and Hayre, L.S.(1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7:307-316.

[11] Fan, W., and Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*, 26(2): 132-139.

[12] Foradori, G. T. (1961). Some non-response sampling theory for two stage designs. *Mimeo*, 297.

[13] Geng, G.Z. and Gao, G. and Ruan, Y.H. and Yu, M.R. and Zhou, Y.H.(2016). Behavioral risk profile of men who have sex with men in beijing, China: results from a cross-sectional survey with randomized response techniques. *Chinese Medical Journal*, 129(5): 523-529.

[14] Gill, T.S. and Tuck, A. and Gupta, S. and Crowe, M. and Figueroa, J. (2013). A field test of optional unrelated question randomized response models: estimates of risky sexual behaviors. *Springer Proceedings in Mathematics and Statistics Series*, 64:135-146.

[15] Greenberg, B.G. and Abul-Ela (1969). The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, 64(326):520-539.

[16] Gregoire, T.G. and Salas, C. (2009). Ratio estimation with measurement error in the auxiliary variate. *Biometrics*, 65:590-598.

[17] Gupta, S., Gupta, B., and Singh, S. (2002). Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and inference*, 100:239-247.

[18] Gupta, S. and Shabbir, J. (2004). Sensitivity estimation for personal interview survey questions. *Statistica*, 64:643-653.

[19] Gupta, S., Shabbir, J., and Sehra, S.J. (2010). Mean and sensitivity estimation in optional randomized response models. *Journal of Statistical Planning and Inference*, 140(10):2870-2874.

[20] Gupta, S., Shabbir, J., Sousa, R., and Corte-Real, P. (2012). Estimation of the mean of a sensitive variable in the presence of auxiliary information.*Communications in Statistics - Theory and Methods*, 41(13-14):2394-2404.

[21] Gupta, S., Kalucha J., Shabbir, J., and Dass, B.K. (2014). Estimation of Finite Population Mean Using Optional RRT Models in the Presence of Nonsensitive Auxiliary Information . *American Journal of Mathematical and Management Sciences*, 33(2):147-159.

[22] Gupta, S., Kalucha,G., Shabbir. J. (2015). A regression estimator for finite population mean of a sensitive variable using an optional randomized response model. *Communications in Statistics - Simulation and Computation*, 46(3):2393–2405.

[23] Gupta, S., Shabbir. J., Sousa, R., Corte-Real, P. (2016). Improved exponential type estimators of the mean of a sensitive variable in the presence of nonsensitve auxiliary information. *Communications in Statistics - Simulation and Computation*, 45(9):3317–3328.

[24] Gupta, S., Mehta, S., Shabbir. J. and Khalil, S. (2018). A unified measure of respondent privacy and model efficiency in quantitative RRT models. *Journal of Statistical Theory and Practice*, 12(3):506-511.

[25] Hansen, M. H., Hurwitz, W. N.(1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41:517–529.

[26] Kadilar C. and Cingi, H. (2003). Ratio estimator in stratified sampling. *Biometrical Journal*, 45:218–225.

[27] Kadilar C. and Cingi, H. (2005). A new estimator in stratified random sampling. *Communication in Statistics-Theory and Methods*, 34:597–602.

[28] Kadilar, C. and Cingi, H. (2005). A new estimator using two auxiliary variables.*Applied Mathematics and Computation*, 162:901-908.

[29] Kadilar, C. and Cingi, H. (2006). Improvement in estimating the populatoin mean in simple random sampling.*Applied Mathematics Letters*, 19(1):75-79.

[30] Kadilar, C., Candan, M., and Cingi, H. (2007). Ratio estimators using robust regression.*Hacettepe Journal of Mathematics and Statistics*, 36(2):181-188.

[31] Kalucha G, Gupta S., Dass B.K. (2015). Ratio estimation of finite population mean using optional randomized response models.*Journal of Statistical Theory and Practice*, 9(3):633-645.

[32] Kerkvliet, J. (1994). Estimating a logit model with randomized data: the case of cocaine use. *Australian and New Zealand Journal of Statistics*, 36:9-20.

[33] Khalil, S., Gupta, S., and Hanif, M. (2018) A generalized estimator for finite population mean in the presence of measurement errors in stratified random sampling. *Journal of Statistical Theory and Practice*, 12(2):311-324.

[34] Khalil, S., Noor-ul-Amin, M., and Hanif, M. (2018). Estimation of population mean for a sensitive variable in the presence of measurement error. *Journal of Statistics and Management Systems*, 21(1):81-91.

[35] Khalil, S., Gupta, S., and Hanif. M. (2018). Estimation of finite population mean in stratified sampling using scrambled responses in the presence of measurement errors. *Communications in Statistics - Theory and Methods*, 48(6):1553-1561.

[36] Khalil, S., Zhang, Q., and Gupta, S. (2019) Mean Estimation of Sensitive Variables under Measurement Errors using Optional RRT Models. *Communications in Statistics – Simulation and Computation,* DOI: 10.1080/03610918.2019.1584298

[37] Khare, B.B. and Srivastava, S. (1993). Estimation of population mean using auxiliary character in presence of non-response. *National Academy Science Letters, India*, 16:111-114.

[38] Khare, B.B. and Srivastava, S. (1995). Study of conventional and alternative two-phase sampling ratio, product and regression estimators in presence of non-response. *Proceedings of the Indian National Science Academy*, 65:195-203.

[39] Khare, B.B. and Srivastava, S. (1997). Transformed ratio type estimators for the population mean in presence of non-response. *Communications in Statistics - Theory and Methods*, 26:1779-1791.

[40] Khare, B.B. and Srivastava, S. (2010). Generalized two phase estimators for the population mean in the presence of non-response. *Aligarh Journal of Statistics*, 30:39-54.

[41] Koyuncu, N., and C. Kadilar. (2008) Ratio and Product estimators in stratified random sampling. *Journal of Statistical planning and Inference*, 139 (8):2552–2558.

[42] Koyuncu, N., and C. Kadilar. (2009) Family of estimators of population mean using two auxiliary variables in stratified random sampling. *Communications in Statistics-Theory and Methods*, 38 (14):2398–2417.

[43] Koyuncu, N., and C. Kadilar. (2010) On the family of estimators of population mean in stratified random sampling. *Pakistan Journal of Statistics* , 26 (2):427–443.

[44] Kerkvliet, J. (2009). Efficient estimators for the population mean. *Hacettepe Journal of Mathematics and Statistics*, 38(2):217-225.

[45] Kumar, M., Singh, R., Sawan, N., and Chauhan, P. (2011). Exponential ratio method of estimation in the presence of measurement errors. *International Journal of Agricultural and Statistics Sciences*, 7(2):457-461.

[46] Kumar, M., Singh, R., Singh, A.K. and Smarandache, F. (2011). Some ratio type estimators under measurement errors. *World Applied Sciences Journal*, 14(2):272-276.

[47] Kumar, S and Bhougal, S. (2011). Estimation of the population mean in presence of non-response. *Communications of the Korean statistical society*, 18(1):1-12.

[48] Manisha and Singh, R.K. (2001). An estimation of population mean in the presence of measurement errors. *Journal of the Indian Society of Agricultural Statistics* 54(1): 13-18.

[49] Millar, M.M., and Dillman, D.A. (2011). Improving response to web and mixed-mode surveys. *Public Opinion Quarterly*, 75(2):249-269.

[50] Murthy, M.N.(1963). Product method of estimation.*Sankhya*, 26:69-74.

[51] Nangsue, N. (2009). Adjusted ratio and regression type estimators for estimation of population mean when some observations are missing. *World Academy of Science, Engineering and Technology*, 53:787-790.

[52] Perri, P.F. (2008). Modified randomized devices for Simmons' model. *Model Assisted Statistics and Applications*, 3:233-239.

[53] Pollock, K. and Bek, Y. (1976). A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association*, 71(356):884-886.

[54] Robson, D.S.(1957). Applications of multivariate polykays of the theory of unbiased ratio-type estimation.*Journal of the American Statistical Association*, 52:511-522.

[55] Salas, C. and Gregoire, T.G. (2010). Statistical analysis of ratio estimators and their estimators of variances when the auxiliary variate is measured with error.*European Journal of Forest Research*, 129:847-861.

[56] Saha, A. (2008). A randomized response technique for quantitative data under unequal probability sampling.*Journal of Statistical Theory and Practice*, 2(4):589-596.

[57] Shabbir J. and Gupta S. (2005). Improved ratio estimators in stratified sampling. *American Journal of Mathematical and Management Sciences*, 25:293–311.

[58] Shabbir J. and Gupta S. (2006). A new estimator of population mean in stratified sampling. *Communication in Statistics Theory and Methods*, 35:1201–1209.

[59] Shabbir, J. and Gupta, S. (2007). On improvement in variance estimation using auxiliary information.*Communications in Statistics-Theory and Methods*, 36(12):2177-2185.

[60] Shabbir, J. and Gupta, S. (2010). Estimation of the finite population mean in two-phase sampling when auxiliary variables are attribute.*Hacettepe Journal of Mathematics and Statistics*, 39(1):121-129.

[61] Shabbir, J. and Khan, N.S. (2013). Some modified exponential ratio-type estimators in the presence of non-response under two-phase sampling scheme.*Electronic Journal of Applied Statistical Analysis*, 6(1):1-17.

[62] Shabbir, S. (1997). Ratio method of estimation in the presence of measurement errors.*Journal of the Indian Society of Agricultural Statistics*, 50(2): 150-155.

[63] Shukla, D., Pathak, S. and Thakur, N.S. (2012). An estimator for mean estimation in presence of measurement error.*Research and Reviews: A Journal of Statistics*, 1(2): 1-8.

[64] Singh, S. and Joarder, A.H. and King, M.L.(1996). Regression analysis using scrambled responses.*Austrian Journal of Statistics*, 38(2):201-211.

[65] Singh, H.P. and Karpe, N. (2007). Effect of measurement errors on a class of estimators of population mean using auxiliary information in sample surveys. *Journal of Statistical Research of Iran*, 4:175-189.

[66] Singh, H.P. and Karpe, N. (2008). Estimation of population variance using auxiliary information in the presence of measurement errors. *Statistics in Transition-New Series*, 9(3): 443-470.

[67] Singh, H.P. and Kumar, S. (2008). Estimation of mean in presence of non-response using two-phase sampling scheme. *Statistical Papers*, 51:559-582.

[68] Singh, H.P. and Karpe, N. (2009). On the estimation of ratio and product of two population means using supplementary information in presence of measurement error. *Statistica*, 1:27-47.

[69] Singh, H.P. and Kumar, S. (2009). A general procedure of estimating the population mean in the presence of non-response under double sampling using auxiliary information. *Statistics and Operations Research Transactions*, 33:71-84.

[70] Singh, H.P. and Karpe, N. (2010). Estimation of mean, ratio and product using auxiliary information in the presence of measurement errors in sample surveys. *Journal of Statistical Theory and Practice*, 4(1): 111-136.

[71] Singh, H.P. and Kumar, S. and Kozak, M. (2010). Improve estimation of finite population mean using sub-sampling to deal with the non-response in two-phase sampling scheme. *Communications in Statistics - Theory and Methods*, 39(5):791-802.

[72] Singh, H. P., and N. Karpe.(2010). Effect of measurement errors on the separate and combined ratio and product estimators in stratified random sampling.*Journal of Modern Applied Statistical Methods*, 9(2):338–402.

[73] Singh, H.P. and Kumar, S. (2011). Combination of regression and ratio estimate in presence of non-response. *Brazilian Journal of Probability and Statistics*, 25(2):205-217.

[74] Singh, H. P., and Karpe, N. (2010). Effect of measurement errors on the separate and combined ratio and product estimators in stratified random sampling. *Journal of Modern Applied Statistical Methods*, 9(2):338–402.

[75] Singh, V.K., Singh, R., and Smarandache, F.(2014). Different-type estimators for estimation of mean in the presence of measurement error. *arXiv preprint arXiv:1410.0279.*

[76] Singh, S.R. and Sharma, P. (2015). Method of estimation in the presence of non-response and measurement errors simultaneously. *Journal of Modern Applied Statistical Methods*, 14(1):107-121.

[77] Sousa R., Shabbir S., Corte-Real P., and Gupta S. (2010). Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. *Journal of Statistical Theory and Practice*, 4(3):495-507.

[78] Sukhatme P. and Sukhatme, B. (1970). Sampling theory of ssurvey with applications. *Iowa State University Press.*

[79] Srinath, K.P. (1971). Multiphase sampling in non-response problems. *Journal of the American Statistical Association*, 66:583-586.

[80] Srivastava, A. and Shalabh, S. (2001). Effect of measurement errors on the regression method of estimation in survey sampling. *J. Statist. Res.*, 35(2):35-44.

[81] Subramani, J. and Kumarapandiyan, G. (2012). Variance estimation using quartiles and their functions of an auxiliary variable. *International Journal of Statistics and Applications*, 2:67-72.

[82] Tarray, T.A., and Singh, H. (2015). A general procedure for estimating the mean of a sensitive variable using auxiliary information. *Investigacion Operacional*, 36(3):268-279.

[83] Turgut, Y. and Cingi, H. (2008). New generalized estimators for the population variance using auxiliary information. *Hacettepe Journal of Mathematics and Statistics*, 41(5):627-636.

[84] Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63-69.

[85] Warner, S.L.(1971). The linear randomized response model. *Journal of the American Statistical Association*, 66(336):884-888.

[86] Wu, J.W., Tian, G.L., and Tang, M.L. (2008). Two new models for survey sampling with sensitive characteristics: Design and analysis. *Metrika*, 67:251-263.

[87] Yan, Z., Wang, J., & Lai, J. (2008) An efficiency and protection degree-based comparison among the quantitative randomized response strategies. *Communications in Statistics-Theory and Methods*, 38(3): 400-408

[88] Zahid, E. and Shabbir, J. (2018). Estimation of population mean in the presence of measurement error and non-response under stratified random sampling. PLoS ONE 13(2): e0191572. *10.1371/journal.pone.0191572*.

[89] Zhang, Q., Kalucha, G., Gupta, S. and Khalil, S. (2018). Ratio Estimation of the Mean under RRT Models. *Journal of Statistics and Managements Systems*, 22(1):97-113.

[90] Zhang, Q. (2016). Ratio Estimation of the Mean under RRT Models. (Master's Thesis) https://libres.uncg.edu/ir/uncg/f/Zhang_uncg_0154M_12079.pdf

[91] Zhang, Q., Khalil, S. and Gupta, S. (2020) Mean Estimation of Sensitive Variables under Non-response and Measurement Errors using Optional RRT Models. (under review)

[92] Zhang, Q., Khalil, S., and Gupta, S. (2020) Mean Estimation of Sensitive Variables under Non-response Measurement Errors using Optional RRT Models in Stratified Random Sampling. (preprint)

APPENDIX A

LIST OF PUBLICATIONS

1. Zhang, Q., Gupta, S., Kalucha, G., and Khalil, S. (2019) Ratio estimation of the mean under RRT models. *Journal of Statistics and Management Systems* 22(1): 97-113.

2. Khalil, S., Zhang, Q., and Gupta, S. (2019) Mean Estimation of Sensitive Variables under Measurement Errors using Optional RRT Models. *Communications in Statistics – Simulation and Computation,* DOI: 10.1080/03610918.2019.1584298

3. Zhang, Q., Khalil, S. and Gupta, S. (2020) Mean Estimation of Sensitive Variables under Non-response and Measurement Errors using Optional RRT Models. (Submitted)

4. Zhang, Q., Khalil, S., and Gupta, S. (2020) Mean Estimation of Sensitive Variables under Non-response Measurement Errors using Optional RRT Models in Stratified Random Sampling. (Preprint)