ON ESTIMATION OF POPULATION PARAMETERS IN SAMPLING THEORY USING A SENSITIVE VARIABLE

THESIS

SUBMITTED TO

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
(A CENTRAL UNIVERSITY)
LUCKNOW

FOR THE AWARD OF DEGREE OF

Doctor of Philosophy

IN

STATISTICS

Submitted by

Tarushree Bari

Enrolment No: 437/14

Co-Supervisor

Dr. Subhash Kumar Yadav
Associate Professor

Supervisor

Dr. Amit Kumar Misra
Assistant Professor

DEPARTMENT OF STATISTICS
SCHOOL OF PHYSICAL & DECISION SCIENCES
BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
(A CENTRAL UNIVERSITY)
Vidya Vihar, Raebareli Road, Lucknow-226025 (U.P.), India

2022

Dedicated

To

My Beloved Parents

DECLARATION

I, **Tarushree Bari**, Enrolment No. 437/14, hereby declare that the work which is being presented in the thesis entitled **"On Estimation of Population Parameters in Sampling Theory using a Sensitive Variable"** in fulfillment of the requirements for the award of the degree of Doctor of Philosophy and submitted in the Department of Statistics, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow (U.P.), India, is an authentic record of my own work carried out during the research period under the supervision of Dr. Amit Kumar Misra, Assistant Professor and co-supervision of Dr. Subhash Kumar Yadav, Associate Professor, Department of Statistics, School of Physical & Decision Sciences, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow (U.P.), India.

The matter presented in this thesis has not been submitted by me for the award of any other degree or diploma to this or any other Institute. I also declare that the thesis is essentially free from all kinds of plagiarism.

Date:
Place:

(Tarushree Bari)
Research Scholar
Department of Statistics,
School of Physical & Decision Sciences,
Babasaheb Bhimrao Ambedkar University,
Lucknow-226025, India.

CERTIFICATE

This is to certify that the thesis titled **"On Estimation of Population Parameters in Sampling Theory using a Sensitive Variable"** submitted by **Ms. Tarushree Bari** is an original research work and has not been previously submitted in part or full for the award of any other degree or diploma to this or any other university.

The thesis submitted to Babasaheb Bhimrao Ambedkar University, Lucknow satisfies all the requirements as stipulated in the *Master of Philosophy (M. Phil.) / Doctor of Philosophy (Ph.D.) regulations* amended in *2017* and it is fit for submission and evaluation for the award of the degree of Doctor of Philosophy of the University.

Co-Supervisor Supervisor

Date: Head of the Department

ACKNOWLEDGMENTS

Throughout the writing of this thesis, I have received a great deal of support and assistance. I take this opportunity to express my sincere gratitude to all of them. The submission of my thesis became possible only with generous help, guidance and sincere support which I recieved at every step from my erudite supervisor Dr. Amit Kumar Misra, Assistant Professor, Department of Statistics, Babasaheb Bhimrao Ambedkar University, Lucknow and co-supervisor Dr. Subhash Kumar Yadav, Associate Professor, Department of Statistics, Babasaheb Bhimrao Ambedkar University, Lucknow. They helped me at every stage from the conception of the problem to the submission of the thesis. My sincere gratitude goes out to them for their kind guidance and support in this academic endeavor.

I want to express my sincerest gratitude to our Head of the Department, Prof. Surinder Kumar, for his continuous encouragement, guidance, and motivation during my research work. Thank him for providing a peaceful and supportive environment in the department.

I am also thankful to the members of Departmental Research Committee for their helpful career advice and valuable suggestions in general and the faculty members of the department Prof. Madhulika Dube, Dr. Rahul Varshney, and Dr. Meenakshi Mishra. Their encouraging and helpful attitude often directly or indirectly benefitted my research.

My deepest thanks to the non-teaching staff cum my dear friend, Mrs. Somya Trivedi for always being there for me. I am also thankful to other non-teaching staff of the department Dr. Amrendra Pratap Bahadur Singh and Mr. Nirmal Singh for their kind cooperation and support. I would like to express my heartfelt appreciation for their assistance in completing the formalities during the course of the present work.

I would like to express my heartwarming thanks to my dear friends Arun and Aditi

and brother Vishwajeet Singh for providing me the determination and patience to continue my hard work especially in the period of doubt and despair. I feel fortunate to express my appreciation to my seniors Dr. Ruby Chanchal, Dr. Vaishali Gupta, Dr. Mradula, Dr. Ankit and my juniours Shivendra, Anuj and Dikshita for their continuous love and support.

I would like to acknowledge my friends Priyam and Agyeya for their moral support and motivation, which drives me to give my best. A special mention of thanks to my dearest friend Sanju for always believing in me and showing the unconditional trust and endless patience.

Finally I owe thanks to a very special person, Vishal for his continuous support and understanding during my pursuit of Ph.D degree that made the completion of my thesis possible. I have no words to express my gratitude and thanks to my parents for their limitless sacrifices to enrich my future whose love and guidance are with me in whatever I pursue.

Last but not least, I conclude my acknowledgment with thanksgiving to great almighty with my both hands for His throughout blessings.

(Tarushree Bari)

LIST OF PUBLISHED/COMMUNICATED RESEARCH PAPERS

1. SUBHASH KUMAR YADAV, AMIT KUMAR MISRA AND TARUSHREE BARI (2022). Searls Estimation Strategy for Population Mean of a Sensitive Study Variable harnessing Non-sensitive Auxiliary Information. *International Journal of Mathematics in Operational Research.* https://doi.10.1504/IJMOR.2021.10045248.

2. SUBHASH KUMAR YADAV, AMIT KUMAR MISRA AND TARUSHREE BARI (2022). Generalized classes of Estimators for Population mean of Sensitive Variable Using Non-sensitive Auxiliary Parameters. *International Journal of Mathematical Modelling and Numerical Optimization.* https://doi.org/10.1504/IJMMNO.2022.10046011.

3. SUBHASH KUMAR YADAV, AMIT KUMAR MISRA AND TARUSHREE BARI (2022). Robust Type Regression Estimator of the Mean of a Sensitive Variable. *Mathematical Population Studies.* Under Revision.

4. SUBHASH KUMAR YADAV, AMIT KUMAR MISRA AND TARUSHREE BARI (2022). Generalized Double Sampling Family of Estimators for Population mean of Sensitive Variable Harnessing Non-sensitive Auxiliary Parameter. *Investigacion Operacional.* Under Revision.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

Chapter 1

Introduction and Literature Review

1.1 Introduction

One of the main purposes of statistical research is to determine the true values of population parameters. However, gathering data from every person of a huge population would be too expensive and time consuming. Rather than conducting a census, we can collect data from a sample and make inferences about the population of interest using sample statistics. A sample may not adequately represent the population due to several sampling or non sampling errors. Sampling error occurs when we operate with a subset of the population rather than the entire population. It is often possible to reduce it by increasing the sample size. While on the contrary, non-sampling errors can be a result of a variety of factors, including respondent error, measurement error, and non-response. As a result, meaningful inferences may be drawn only if the sample accurately represents the population, otherwise, the sample is skewed, and the conclusions drawn from the study are not reliable. Based on the circumstances, we could employ a variety of sampling procedures such as

simple random sampling, cluster sampling, and stratified random sampling, in order to obtain a representative sample. We mostly add information on a variable that is highly linked with the study variable. This additional information is referred to as auxiliary, ancillary, or previous information, and the variable from which it is collected is referred to as auxiliary or ancillary variable. The information on auxiliary variable may be known in advance, based on previous data, a pilot survey, or the observer's experience. When a sample is drawn, one can use auxiliary information to enhance the accuracy of estimation, regardless of the sampling design used. The ratio and regression techniques of sampling include auxiliary information into the estimation procedure to increase the precision of estimates, i.e., to provide estimates that are close to the corresponding population values, while also increasing the estimator's efficiency. It has been statistically demonstrated that the incorporation of auxiliary information in probability sampling reduces the Mean Square Error (MSE) of the estimator of the population parameter by a significant amount. However, it is highly dependent on the manner in which the estimator has been proposed, that is, on the way that the role of auxiliary information has been taken into consideration.

As discussed earlier, a competent sampling technique can assess, whether a sample is actually representative of population or not. Two sample techniques are presented below to discuss this issue.

1.1.1 Simple Random Sampling

Simple random sampling is a sampling strategy that uses an unbiased selection process to ensure that each element of the population has an equal probability of being chosen. Each individual in the sample is assigned a number, and then a sample is selected at random. The simple random sampling method is amongst the most basic and widely used methods of data collection, since it is intended to produce an unbiased representative of a population. The whole data set is represented by a random subset of selected individuals. For example, to select a simple random sample of 20 employees from an organisation, we may assign a number to each employee and then select a sample by randomly generated 20 numbers.

1.1.2 Double Sampling

In double sampling, first a sample of units is chosen for the purpose of gathering auxiliary information first, and then a second sample is chosen for the purpose of observing the variable of interest in additional to the auxiliary information. It is beneficial when collecting data on auxiliary variable X, is significantly cheaper and faster than collecting data on the study variable Y, and the correlation between X and Y is considerably high. Initially, a random sample of size n' is drawn from a population of size N, followed by a random sample of size n from the first sample of size n'. For example, conducting forest surveys in remote areas is a tedious and expensive task. However, gathering data from satellite is comparatively cheaper and the species are strongly correlated to the terrestrial region. We take double sample with aerial photographs in the first phase and field plots in the second

phase.

The social desirability biasness is one of the primary concerns in survey method addressing sensitive questions. In survey research, major topics of interest contain capricious questions that respondents find challenging to respond. To gather personal information on sensitive topics like drug addictions, abortion, crime, taxations, support of terrorism is a tedious task. For Example

(a) Are you indulge in terrorism funding?

(b) Do you smoke regularly?

(c) Have you disclosed the real income in ITR filing?

(d) Are you involved in child trafficking?

(e) Have you gone through any mental harassment?

Respondents feel insecure to present themselves in particular community as some have fear of being exposed while others find it derogative to reveal the truth. In order to feel socially accepted in the society, respondents might give delusive reactions. To solve this issue, Warner (1965) gave the randomized response technique model to eliminate the bias in the response by dividing the population into two groups. The respondent is asked to say 'Yes' or 'No' depending upon the specified randomized design. For estimating population mean of the sensitive variable, we often use the information of the subsidiary variable which attains non-sensitive characteristics in nature and is highly correlated with the sensitive variable.

The primary goal of this thesis is to strengthen the parameter estimation of a sensitive variable by making use of non-sensitive auxiliary information in the estimation process. In order to accomplish this, we introduce some population mean estimators that are based on the additive randomized response technique model. Based on the data, it is possible to derive expression for the bias and mean square error for each of the proposed estimators. Additionally, for each of the study estimators, a comprehensive simulation study is conducted, followed by an application to a real-world dataset. All of the application domains in this collection are developed using the statistical software R.

Now, we provide some useful definitions and notation in the following sections, that are used in the thesis.

1.2 Randomized Response Technique

Randomized Response Technique (RRT) is a methodology for gathering sensitive information from persons in which survey interviewers and data processors are ignorant of which of two alternative the respondents actually replied. RRT models can be broadly classified into three types: Full RRT models, in which all respondents provide scrambled responses; Partial RRT models, in which a randomly selected sub-group of respondents is asked to provide truthful responses; and Optional RRT models, in which the respondent has the option of providing either a scrambled response or a truthful response, depending on whether the respondent considers the question sensitive or non-sensitive.

1.2.1 Warner's RRT Model (1965)

When asking questions about sensitive behavioural patterns, Warner (1965) proposed the randomized response technique as a structured survey technique to minimize potential bias due to non-response and social norms. This technique was later adopted by other researchers. Respondents are instructed to use a randomization device, such as a coin, a deck of cards, or spinners, the outcome of which is unknown to the enumerator, in order to complete the survey. Consider that every individual in a total population corresponds to either Group A or B, and it is needed to survey the proportion of people in Group A. We select n people at random, with replacement from the population, to interview. Each interviewer is given a spinner that has been marked so that it points to the letter A with probability p, and to the letter B with probability $(1-p)$, $0 < p < 1$. The interviewee is then required to spin the spinner in front of the interviewer and disclose unless the spinner indicates to the letter identifying the group they belong to. In other words, the interviewee is simply needed to respond affirmatively or negatively depending on whether or not the spinner points to the correct group, he/she is not obligated to report whatever group the spinner has pointed. Assuming these yes/no reports are accurate, maximum probability estimations of the underlying population proportion are obtained. Let us consider

$$\pi = \text{chance of occurring A in the population}$$

$$p = \text{the probability of the spinner hitting at A}$$

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ sample element says yes,} \\ 0, & \text{if the } i^{th} \text{ sample element says no.} \end{cases}$$

Then

$$P(X_i = 1) = \pi p + (1-\pi)(1-p),$$

$$P(X_i = 0) = (1-\pi)p + \pi(1-p).$$

Also ordering the sample such that first n_1 report "yes" while the second $(n-n_1)$ report "no". The estimate of π is given by

$$\pi = \frac{(p-1)}{(2p-1)} + \frac{n_1}{(2p-1)n}.$$

The respective model is given by

$$Y = Z\pi + u,$$

assuming $\quad E(u) = 0 \quad$ and $\quad E(uu') = \phi,$

where ϕ is matrix of fixed constants and Z is a matrix of constant with rank p.

1.2.2 Greenberg's RRT Model (1971)

When it comes to binary data, Warner (1965) discussed the above randomization device. Greenberg *et al.* (1971) model is an extension of the binary model given by Warner (1965). The response obtained through sensitive inquiry has a probability density function g whereas the response collected through non-sensitive inquiry has the probability density function h. Answers to sensitive questions are given with probability p, while non-sensitive questions are given with probability $(1-p)$. As a result, the observed response has a distribution that may be approximated as

$$\pi = pg(z) + (1-p)h(z). \tag{1.2.1}$$

1.2.3 The Additive Model

Pollock and Bek (1976) suggested an additive model in which a sensitive property Y and a random value S from a known distribution are combined, and the respondent is required to add the results. The observed response indicated by Z, is given by

$$Z = Y + S.$$

The respective mean and variance of the variable Z, are given by

$$\mu_z = \mu_y + \mu_s \quad \text{and} \quad \sigma_z^2 = \sigma_y^2 + \sigma_s^2. \tag{1.2.2}$$

1.2.4 The Multiplicative Model

Eichhron and Hayre (1983) suggested multiplicative approach to handle issues related to sensitive data. Under this model, the respondent is required to multiply his/her sensitive value Y by a random value S, drawn from a known distribution in this model. Thus, the resultant response is given by

$$Z = YS, \tag{1.2.3}$$

with the respective mean and variance, given by

$$\mu_z = \mu_y\, \mu_s \quad \text{and} \quad \sigma_z = \mu_s^2\sigma_y^2 + \mu_y\sigma_s^2 + \sigma_y^2\sigma_s^2. \tag{1.2.4}$$

1.2.5 Gupta *et al.* ORRT Model (2002)

Gupta *et al.* (2002) examined an alternative randomized response model, i.e., Optional Randomized Response Model (ORRT) in which the respondent is given the option of providing a genuine or scrambled response. The ORRT is premised on the basis that inquiry might be sensitive to one respondent but not to other. Hence, respondents submit the scrambled response depending upon whether they find the question sensitive or not. However, because of the confidentiality, the interviewer will not be able to tell which type of response, the interviewee will provide. Sensitivity levels are also estimated in addition to the mean and variation of the variable that is being studied. The proposed model consist of

$$Z = S^G X, \tag{1.2.5}$$

where G is a random variable that takes the value as

$$G = \begin{cases} 1, & \text{if the respose is scrambled,} \\ 0, & \text{otherwise.} \end{cases}$$

1.2.6 Gupta *et al.* ORRT Model (2010)

Gupta *et al.* (2010) used a split-sample strategy to suggest an additive ORRT model by separating the sample into two divisions. Both the subgroups give scrambled response by using two different scrambled response S_1 and S_2. The reported response Z_i, $i = 1,2$, for

each i^{th} subgroups, is given by

$$Z_i = \begin{cases} Y, & \text{with the probability } 1-W, \\ Y+S_i, & \text{with the probability } W. \end{cases}$$

1.2.7 Diana and Perri's Linear RRT Model (2011)

The purpose of using an RRT model is to protect the privacy of respondents. In this context, a mixture of additive and multiplicative techniques can increase respondents' confidence in their privacy protection because two scrambling variables will be incorporated into the model. Let T and S be two scrambling variables independent of each other with mean μ_T and μ_S and variances σ_T^2 and σ_S^2. Both T and S are not related to the study variable Y. Diana and Perri (2011) developed a more broad linear combination model, which is represented by

$$Z = TY + S. \tag{1.2.6}$$

Assuming the value of $\mu_T = 1$ and $\mu_S = 0$, the mean and variance of Z, are given by

$$E(Z) = \mu_Y \quad \text{and} \quad V(Z) = \sigma_S^2(\mu_Y^2 + \sigma_Y^2) + \sigma_Y^2 + \sigma_T^2. \tag{1.2.7}$$

1.3 Robust Regression

Using the Ordinary Least Squares (OLS) method in regression analysis would not be appropriate when attempting to solve a problem that contains outlier or extreme observations. We require a method of robust estimation in which the value of the estimator is not significantly affected by outlier or extreme observations. Many theoretical efforts have been

made since 1960, to construct statistical processes that are resistant to minor deviations from assumptions, i.e., robust in the case of outliers and stable in the presence of small departures from the presumed parametric model. In reality, it is well-known that classical optimal techniques fail when the strict assumptions of a model are violated. It is also well understood that screening the data, removing outliers, and then applying traditional inferential processes is not a simple or effective method of proceeding. To begin with, identifying outliers or influential observations in multivariate or highly structured data can be challenging. Also, rather than removing an observation, it may be desirable to reduce the weight of uncertain observations. Furthermore, removing outliers reduces the sample size that may have an impact on the distribution theory, and the variances from the cleaned data may be underestimated. Therefore empirical data suggests that effective robust techniques outperform strategies based on outlier rejection.

In order to determine a regression model, we discuss M estimation, S estimation, and MM type robust regression estimation in this thesis. M estimation is considered a robust extension of the maximum likelihood approach, whereas S estimation and MM estimation are extensions of the M estimation method.

1.4 Notation

Let Y be the research variable, which comprises delicate features that may not be accessed exactly as a result of the respondent's response. Let X be a non-sensitive secondary variate that is correlated positively to Y. Let S be a scrambled variate with a predefined distribution

which is uncorrelated with Y and X. The sensitive variable Y is scrambled using the variable S. To assess Y, each respondent is instructed to choose a random number from the S distribution, say s, and add it to the actual value of Y. Let Z denote the reported scrambled response to Y that was initially suggested by Warner (1965) and later elaborated by Pollock and Bek (1976), given by

$$Z = Y + S.$$

- N : Size of the population
- n : Size of the sample
- $\overline{Y}, \overline{X}$: Population means of Y and X, respectively
- $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$: Sample means of Y and X, respectively
- ρ_{yx} : Correlation Coefficient between Y and X
- $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2$: Population mean squares of the sensitive variable Y
- $S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$: Population mean squares of the non-sensitive auxiliary variable X
- $S_z^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)^2$: Population mean squares of the observed response Z
- S_{zx} : Covariance between Z and X
- C_y, C_x : Coefficients of Variations of Y and X, respectively
- MSE$(\cdot)$: Mean squared error of the estimator

- $f = \frac{n}{N}$: Sampling fraction

Let a simple random sample of size n be drawn without replacement from a finite population $U = (U_1, U_2, \ldots, U_N)$ in view to estimate the mean of the sensitive variable. For the i^{th} unit, let y_i and x_i respectively, be the sample values of the study variable and the auxiliary variable.

A brief outline of the literature is presented in the following section.

1.5 Literature review

Many survey practitioners have introduced and improved ratio, product, difference, exponential, and linear regression estimators in their search for the most accurate and efficient estimator of population mean. Cochran (1940, 1942) recommended the use of ratio estimator if the study and auxiliary variable are closely related and have a positive correlation.

Murthy (1967) suggested product estimators when the research characteristic and auxiliary variable have a negative correlation. However, the ratio method has a more extensive range of applications because of its concise computational approach to research methodology. To obtain a more accurate estimate, several authors have incorporated the prior information of one or more population parameters. Searls (1964) employed the coefficient of variation at the estimation stage and developed more precise estimators of population mean. Following Searls (1964) methodology, Sen (1978), Upadhyaya and Singh (1984) and Khoshnevisan *et al.* (2007) utilized coefficient of variation of auxiliary variable

to estimate population mean of the study variable.

An extensive literature review has been carried out on mean estimation when the primary variable is highly sensitive and there are no auxiliary variables. Some examples are Eichhron and Hayre (1983), Gupta and Shabbir (2004), Gupta *et al.* (2002, 2010), Saha (2008), and Perri (2008).

Later, a number of researchers including Sousa *et al.* (2010), Gupta *et al.* (2012, 2017), Koyuncu *et al.* (2014), Kalucha *et al.* (2015) attempted to estimate the population mean of the sensitive study variable in presence of non-sensitive study variable. Some well-known population mean estimators of sensitive study variable are provided below.

1.5.1 Ratio Estimator

Using the idea of Pollock and Bek (1976), Sousa *et al.* (2010) suggested a ratio estimator, in which the mean of Y is estimated using the RRT model, which is further enhanced by the addition of an auxiliary variable X. Using the available non-sensitive auxiliary information, Sousa *et al.* (2010) improved the RRT estimator of the population mean of the sensitive variable by introducing ratio estimator given by

$$t_R = \overline{z}\left(\frac{\overline{X}}{\overline{x}}\right). \tag{1.5.1}$$

This estimator is biased for the population mean of the sensitive study variable. The bias and MSE of the aforementioned estimator, accurate up to the first degree of approxi-

mation are given by

$$\text{Bias}(t_R) = \lambda \overline{Z}(C_x^2 - \rho_{zx} C_z C_x), \text{ and}$$

$$\text{MSE}(t_R) = \lambda \overline{Z}^2 (C_z^2 + C_x^2 - 2\rho_{zx} C_z C_x).$$

Sousa *et al.* (2010) further modified the ratio estimator by making use of some known auxiliary parameters, given by

$$t_{TR} = \overline{z} \left(\frac{c\overline{X} + d}{c\overline{x} + d} \right), \tag{1.5.2}$$

where c and d are auxiliary parameters such as the coefficient of skewness, coefficient of kurtosis, etc. The bias and MSE of the modified ratio estimator are given by

$$\text{Bias}(t_{TR}) = \lambda \overline{Z}(\phi^2 C_x^2 - \phi \rho_{zx} C_z C_x), \text{ and}$$

$$\text{MSE}(t_{TR}) = \lambda \overline{Z}^2 (C_z^2 + \phi^2 C_x^2 - 2\phi \rho_{zx} C_z C_x),$$

where $\phi = \frac{c\overline{X}}{c\overline{X}+d}$.

1.5.2 Ordinary Regression Estimator

Gupta *et al.* (2012) provided a regression estimator that performs better than the ratio estimator even when the primary and auxiliary variables have a low correlation. The basic premise is that the primary variable is sensitive, but there is a non-sensitive auxiliary variable that is positively related to the primary variable. The typical regression estimator for

the finite population mean of the susceptible study variable Y is as follows

$$t_{Reg} = \bar{z} + b_{zx}(\overline{X} - \bar{x}), \tag{1.5.3}$$

where $b_{zx} = \dfrac{s_{zx}}{s_x^2}$ is the sample regression coefficient of Z on X. Up to the first order of approximation, the bias of this regression estimator is given as

$$\text{Bias}(t_{Reg}) = -\lambda \beta_{zx} \left(\frac{\mu_{12}}{\mu_{11}} - \frac{\mu_{03}}{\mu_{02}} \right),$$

where $\beta_{zx} = \frac{S_{zx}}{S_x^2} = \frac{S_{yx}}{S_x^2} = \rho_{yx} \frac{S_y}{S_x} = \beta_{yx}$ is the population regression coefficient and $u_{rs} = E(z_i - \overline{Z})^r (x_i - \overline{X})^s$. The mean square error of the regression estimator, up to the first order of approximation, is given as

$$\begin{aligned} \text{MSE}(t_{Reg}) =& \lambda \overline{Z}^2 C_Z^2 (1 - \rho_{zx}^2) \\ =& \lambda S_y^2 \left[\left(1 + \frac{S_s^2}{S_y^2} \right) - \rho_{zx}^2 \right]. \end{aligned}$$

1.5.3 Generalized Exponential Estimator

Koyuncu *et al.* (2014) examined the use of exponential-type estimators in order to provide more efficient estimation of the mean of a sensitive variable with the help of non-senitive auxiliary information. The model operates according to the same presumptions as the generalized regression-cum ratio estimator and the ordinary regression estimator, and is given by

$$t_{GE} = [\omega_1 \bar{z} + \omega_2 (\overline{X} - \bar{x})] \exp \left(\frac{\overline{X} - \bar{x}}{\overline{X} + \bar{x}} \right), \tag{1.5.4}$$

where ω_1 and ω_2 are the suitable constants. Considering the first order of approximation, the bias of this generalized exponential estimator is provided as follows

$$\text{Bias}(t_{GE}) \approx (\omega_1 - 1)\overline{Z} + \lambda\omega_2\overline{Z}\left(\frac{3}{8}C_x^2 - \frac{1}{2}\rho_{zx}C_zC_x\right) + \frac{1}{2}\omega_2\lambda\overline{X}C_x^2.$$

Considering the first order of approximation, the minimum mean square error of t_{GE}, for the optimum value of ω_1 and ω_2, i.e.,

$$\omega_1 = \frac{1 - \frac{1}{8}\lambda C_x^2}{1 + \lambda C_z^2(1-\rho_{zx}^2)} \quad \text{and} \quad \omega_2 = \frac{\overline{Z}}{\overline{X}}\left[\frac{1}{2} - \omega_1\left(1 - \frac{C_z}{C_x}\rho_{zx}\right)\right],$$

is given as

$$\begin{aligned}\text{MSE}(t_{GE}) &= \overline{Z}^2\left[\left(1 - \frac{1}{4}\lambda C_x^2\right) - \frac{(1-\frac{1}{8}\lambda C_x^2)^2}{1+\lambda C_z^2(1-\rho_{zx}^2)}\right] \\ &= \left[\frac{\text{MSE}(t_{\text{reg}})}{\left(1 + \frac{\text{MSE}(t_{\text{reg}})}{\overline{Z}^2}\right)} - \frac{\lambda C_x^2\{\text{MSE}(t_{\text{reg}}) + \lambda\frac{1}{16}C_x^2\overline{Z}^2\}}{4\left(1 + \frac{\text{MSE}(t_{\text{reg}})}{\overline{Z}^2}\right)}\right].\end{aligned}$$

1.5.4 Exponential Ratio Type Estimator

Following the estimator proposed by Bahl and Tuteja (1991), Gupta *et al.* (2017) suggested the exponential ratio type estimator to estimate population mean of sensitive study variable, given by

$$t_{ER} = \overline{z}\exp\left(\frac{\overline{X} - \overline{x}}{\overline{X} + \overline{x}}\right). \tag{1.5.5}$$

The exponential ratio type estimator's bias and mean square error are provided by

$$\text{Bias}(t_{ER}) = \lambda\overline{Z}\frac{1}{2}\left(\frac{3}{4}C_x^2 - \rho_{zx}C_zC_x\right) \text{ and}$$

$$\text{MSE}(t_{ER}) = \lambda \overline{Z}^2 \frac{1}{4} \left(4C_z^2 - 4\rho_{zx} C_z C_x + C_x^2\right).$$

In a design-based approach, Shahzad *et al.* (2019) and Sanaullah *et al.* (2019) presented a novel class of ratio-type estimators that are more accurate than the existing ones. The class is originally described with the assumption that the research variable has a non-sensitive character, which means that it deals with themes that do not cause embarrassment when respondents are explicitly questioned about them. Singh *et al.* (2020a, 2020b) presented a randomized mechanism that employs blank cards in the randomization technique to improve estimation methodologies of population mean, connected to quantitative sensitive character.

Modifying the work done by Zaman and Bulut (2019), Ali *et al.* (2021) established a more general class of robust-ratio-type estimators for the sensitive setup. Following that, they also suggested a new family of robust-regression-type estimators in simple random sampling schemes. Tiwari and Mehta (2017) and Onyango *et al.* (2022) attempted to reduce the ambiguity of the responses by considering the three-stage RRT models.

In the presence of measurement error, Waseem *et al.* (2021) and Zhang *et al.* (2021) improved the estimate of the population mean for the sensitive variable using simple and stratified random sampling. Auxiliary information and auxiliary variable ranks were taken into account in the study.

Anas *et al.* (2021) proposed a unique class of estimators by incorporating L-moment characteristics into existing estimators. By integrating robust Minimal Covariance Deter-

minant (MCD), Shahzad *et al.* (2021) developed new regression-type ratio estimators for population mean. Gupta *et al.* (2022) presented an Optional Enhanced Trust (OET) quantitative RRT model that reduces the impact of respondents' lack of confidence on additive model by enabling respondents to choose different scrambling approach.

1.6 Outline of the thesis

In Chapter 2, we present a searls type regression estimator for elevated estimation of the population mean of a sensitive study variable in presence of known non-sensitive additional variable under Simple Random Sampling Scheme. The expressions for the bias and the MSE are obtained using the first order of approximation.

In Chapter 3, we generalize the Sousa *et al.* (2010) family of estimators using some new population parameters of auxiliary information based on a RRT. Further, we introduce a new efficient family of estimators for estimating the population mean of sensitivity variable using the approach given in Sousa *et al.* (2010) in the presence of the auxiliary information. The optimal value of Searl's constant is obtained using Lagrange's method of maxima-minima. Theoretical results are supported with a numerical illustration based on real data sets. In addition, a simulation study is carried out to compare the performances of the suggested and competing families of estimators.

Chapter 4 presents a new class of regression type estimator using different robust techniques like Least Trimmed Square Method (LTS), Least Median of Square (LMS), Least Absolute Deviation (LAD), Tukey M, Hampel M, Huber M and Huber MM. The

study also includes a new robust method, that is, the S method of estimation. The theory is supported by real and simulated data.

In Chapter 5, under a two-phase sampling method, we introduce an enhanced double sampling generalize type estimator for the population mean of a sensitive research variable utilizing information from a non-sensitive supplementary variate. Some special cases of the suggested family of estimators are also discussed. The expressions for the bias and mean squared error of the proposed generalized estimators are derived and theoretical comparisons are made with competing estimators. Theoretical conclusions are supported with a numerical illustration based on real-world data. In addition, a simulation analysis is conducted to compare the efficiencies of the suggested and competing family of estimators.

Chapter 6 provides a broad discussion of the major findings and conclusions, as well as some suggestions for further research.

Chapter 2

Searls Estimation Strategy for Population Mean of a Sensitive Study Variable Harnessing Non-Sensitive Auxiliary Information

2.1 Introduction

As mentioned in Chapter 1, Searls (1964) technique reduces the mean square error of the estimator with known values of coefficient of variation and prior knowledge of auxiliary variables. In order to improve estimators with higher precision, researchers design the experiments and sample size in accordance with the preliminary information. The enhanced predictive estimators under the Searls (1964) method seem to be more effective than their corresponding traditional estimators. In other words, the Searls (1964) technique improves the efficiencies of predictive estimators more effectively than the comparable traditional estimators.

In this Chapter, to construct the most efficient estimator for improved estimation of a population mean of a sensitive variable, we propose a regression estimator using Searls (1964) methodology for estimating the population mean of a sensitive study variable when a non-sensitive secondary variable is available. The expressions for the bias and the mean square error are derived utilizing the approximation of order one. The chapter has been presented in different sections.

The review of the existing competing estimators has been discussed in Section 2.2. The proposed estimator and its sampling properties have been studied in Section 2.3. Under the optimum values of the characterizing scalars, the minimum mean square error of the suggested estimator is also provided. Section 2.4 represents the theoretical efficiency contrast of proposed estimator with the competing estimators and the efficiency conditions over competing estimators are obtained. The theoretical efficiency conditions obtained in Section 2.4 are verified through the numerical example in Section 2.5. These results are also verified through the simulated data set in Section 2.6. Results and discussions are given in Section 2.7. Further, the conclusions are discussed in Section 2.8.

Let Y be the variable under consideration that contain fragile characteristics and may not be accessed precisely as a result of response given by the respondent. Let X be a non-sensitive subsidiary variant that is correlated positively with Y. Let S be a scrambled variant whose distribution is known and is uncorrelated with Y and X. The interviewee addresses a scattered reaction for Y, considering the additive model $Z = Y + S$, but reveals a reliable response for X.

Let a simple random sample of size n be drawn without replacement from a finite population $U = (U_1, U_2, \ldots, U_N)$ in view to estimate the population mean of the sensitive variable. For the i^{th} unit, let y_i and x_i respectively, be the sample values of the study variable and auxiliary variable. The scramble variable is assumed to follow normal distribution with mean zero and variance S_s^2. Therefore, $E(Z) = E(Y)$ and $C_z^2 = C_y^2 + \frac{S_s^2}{\overline{Y}^2}$. We derive the bias and MSE of the proposed estimator using the following relative error terms

$$e_0 = \frac{(\overline{z} - \overline{Z})}{\overline{Z}} \quad \text{and} \quad e_1 = \frac{(\overline{x} - \overline{X})}{\overline{X}},$$

such that $E(e_0) = E(e_1) = 0, \quad E(e_0^2) = \lambda C_z^2, \quad E(e_1^2) = \lambda C_x^2 \quad \text{and} \quad E(e_0 e_1) = \lambda \rho_{zx} C_z C_x,$

where $\lambda = (1 - f)/n$.

2.2 Review of Existing Estimators

When the characteristic available on secondary variable X is not taken into consideration, an unbiased estimator of sensitive variable is the usual RRT sample mean, stated as

$$t_0 = \overline{z}.$$

The mean square error of t_0 is given by

$$\text{MSE}(t_0) = \lambda (S_y^2 + S_s^2). \tag{2.2.1}$$

Sousa *et al.* (2010) improved the ordinary RRT estimator by utilizing the available

non-sensitive auxiliary information X and proposed a ratio type estimator, given by

$$t_R = \overline{z}\left(\frac{\overline{X}}{\overline{x}}\right).$$

This estimator is biased for the population mean of the study variable, and the bias and mean square error of the above estimator, correct up to the first order of approximation is given by

$$\text{Bais}(t_R) = \overline{Z}\lambda(C_x^2 - \rho_{zx}C_zC_x),$$

$$\text{MSE}(t_R) = \lambda\overline{Z}^2(C_x^2 - 2\rho_{zx}C_zC_x + C_z^2). \tag{2.2.2}$$

It is noticeable that the use of secondary variate reduces the mean square error by a considerable amount. The reduction is significant when there exists a high correlation between sensitive variable and secondary variable. Gupta *et al.* (2012) proposed an ordinary regression type estimator of the population mean of the sensitive variable Y, given by

$$t_{Reg} = \overline{z} + \hat{\beta}_{zx}(\overline{X} - \overline{x}),$$

where $\hat{\beta}_{zx} = \frac{s_{zx}}{sx^2} = \frac{s_{yx}}{sx^2} = \rho_{yx}\frac{Sy}{Sx} = \beta_{yx}$ represents sample regression coefficient between Z and X. $\rho_{zx} = \frac{\rho_{yx}}{\sqrt{1+\frac{S_x^2}{S_y^2}}}$ is the correlation coefficient between Z and X. The bias and mean square error of the regression estimator using the first order of approximation is given by

$$\text{Bias}(t_{Reg}) \approx -\lambda\beta_{zx}\left(\frac{\mu_{12}}{\mu_{11}} - \frac{\mu_{03}}{\mu_{02}}\right),$$

$$\text{MSE}(t_{Reg}) \approx \lambda\overline{Z}^2C_z^2(1-\rho_{zx}^2). \tag{2.2.3}$$

Gupta *et al.* (2017) proposed the exponential ratio type estimator to approximate the mean of the sensitive variate considering non-sensitive secondary information, given by

$$t_{ER} = \bar{z} \exp\left(\frac{\overline{X} - \bar{x}}{\overline{X} + \bar{x}}\right).$$

The bias and mean square error of the above mentioned estimator using the approximation of order one is given as

$$\begin{aligned} \text{Bias}(t_{ER}) &= \lambda \overline{Z} \frac{1}{2}\left(\frac{3}{4}C_x^2 - \rho_{zx} C_z C_x\right), \\ \text{MSE}(t_{ER}) &= \lambda \overline{Z}^2 \frac{1}{4}\left(4C_z^2 - 4\rho_{zx} C_z C_x + C_x^2\right). \end{aligned} \tag{2.2.4}$$

2.3 Proposed Estimator

Gupta *et al.* (2012) presented a regression estimator that incorporated an auxiliary variable, X to enhance the RRT estimate of the population mean of the sensitive variable. Motivated from Gupta *et al.* (2012) and using Searls (1964) methodology, which provided a biased sample mean with lesser MSE than the usual sample mean, we propose a new searls type regression estimator to estimate the population mean of the sensitive variable which performs better than the above existing literature, given by

$$t_S = k[\bar{z} + b_{zx}(\overline{X} - \bar{x})], \tag{2.3.1}$$

where k is taken as a constant which is to be obtained such that the MSE of the estimator is least and b_{zx} is the sample regression coefficient between Z and X. We obtain the bias

and MSE of the proposed estimator using the Taylor series expansion. On expressing t_S in terms of e's, then multiplying its terms and preserving the terms up to the first order of approximation, we get

$$t_S - \overline{Z} = \overline{Z}\left[k(1+e_0) - k\beta\left(\frac{\overline{X}}{\overline{Z}}\right)e_1 - 1\right] . \tag{2.3.2}$$

Applying expectations on both sides of Equation (2.3.2), we may obtain the bias of t_S as

$$\begin{aligned} E(t_S - \overline{Z}) &= E\left[\overline{Z}\left\{k(1+e_0) - k\beta\left(\frac{\overline{X}}{\overline{Z}}\right)e_1 - 1\right\}\right] \\ \text{Bias}(t_S) &= \overline{Z}(k-1). \end{aligned} \tag{2.3.3}$$

Squaring both the sides of Equation (2.3.2) and considering terms up to the first order of approximation, we get

$$\begin{aligned} \left(t_S - \overline{Z}\right)^2 &= \overline{Z}^2\left[k(1+e_0) - k\beta\left(\frac{\overline{X}}{\overline{Z}}\right)e_1 - 1\right]^2 \\ &= \overline{Z}^2\left[1 + k^2\left\{(1+e_0)^2 + \beta^2\left(\frac{\overline{X}}{\overline{Z}}\right)^2 e_1^2 + 1 - 2\beta(1+e_0)\left(\frac{\overline{X}}{\overline{Z}}\right)e_1\right\}\right. \\ &\quad \left. -2k\left\{(1+e_0) + 2\beta k\left(\frac{\overline{X}}{\overline{Z}}\right)e_1\right\}\right]. \end{aligned} \tag{2.3.4}$$

Taking expectations on both the sides of above Equation (2.3.4), we get the MSE of t_S, as

$$\text{MSE}(t_S) = \overline{Z}^2\left[1 + k^2\left\{1 + \lambda C_z^2 + \beta^2\left(\frac{\overline{X}}{\overline{Z}}\right)^2 \lambda C_x^2 - 2\beta\left(\frac{\overline{X}}{\overline{Z}}\right)\lambda\rho_{zx}C_zC_x\right\} - 2k\right] \tag{2.3.5}$$

Taking partial derivative of the above Equation (2.3.5) to get the optimum value of k, that minimizes the MSE of t_S, we get

$$k = \frac{1}{Q} \quad \text{where,} \quad Q = \left\{ 1 + \lambda C_z^2 + \beta^2 \left(\frac{\overline{X}}{\overline{\overline{Z}}} \right)^2 \lambda C_x^2 - 2\beta \left(\frac{\overline{X}}{\overline{\overline{Z}}} \right) \lambda \rho_{zx} C_z C_x \right\}.$$

Substituting the value of k in (2.3.5), the minimum MSE of the proposed estimator is given by

$$\begin{aligned} \text{MSE}_{\min}(t_S) &= \overline{\overline{Z}}^2 \left(1 + \frac{1}{Q^2} Q - \frac{2}{Q} \right) \\ &= \overline{\overline{Z}}^2 \left(1 - \frac{1}{Q} \right). \end{aligned} \tag{2.3.6}$$

It is also worth mentioning that at $k = 1$, the proposed estimator reduces to regression type estimator (t_{Reg}) given by Gupta *et al.* (2012).

2.4 Efficiency Comparison

In the following section, the suggested estimator is compared with ordinary RRT, ratio, regression and other estimators given in the literature. Under the below four conditions, the proposed estimator provides greater efficiency than all the estimators taken into consideration.

(i) Using Equations (2.2.1) and (2.3.6), we observe

$$\text{MSE}_{\min}(t_S) < \text{MSE}(t_0) \quad \text{if,} \quad \lambda C_z^2 - \left(1 - \frac{1}{Q} \right) > 0.$$

(ii) Using Equations (2.2.2) and (2.3.6), we observe

$$\text{MSE}_{\min}(t_S) < \text{MSE}(t_R) \quad \text{if,} \quad \lambda(C_x^2 - 2\rho_{zx}) - \left(1 - \frac{1}{Q}\right) > 0.$$

(iii) Using Equations (2.2.3) and (2.3.6), we observe

$$\text{MSE}_{\min}(t_S) < \text{MSE}(t_{Reg}) \quad \text{if,} \quad \lambda C_z^2(1 - \rho_{zx}^2) - \left(1 - \frac{1}{Q}\right) > 0.$$

(iv) Using Equations (2.2.4) and (2.3.6), we observe

$$\text{MSE}_{\min}(t_S) < \text{MSE}(t_{ER}) \quad \text{if,} \quad \lambda \frac{1}{4}\overline{Z}^2 \left(4C_z^2 - 4\rho_{zx}C_zC_x + C_x^2\right) - \left(1 - \frac{1}{Q}\right) > 0.$$

2.5 Numerical Illustration

To support the theoretical results, we considered two real populations. Population 1 is taken from Singh (2003), where X denotes the quantity of non-real estate farm loans during the year 1977 and Y denotes the quantity of real estate farm loans during the year 1977. Population 2 is taken from Murthy (1967), where X denotes the area under wheat in the region during 1971 and Y denotes the area under wheat under the region during 1974. The parametric values of both the populations under consideration are provided in the Table 2.1 given below.

To compare the efficiencies of the proposed estimator, we assessed the PRE of all the competing estimators under consideration for both populations. PRE for various estima-

Table 2.1: Descriptive Statistics of the Real Populations

S. No.	Information	Population 1	Population 2
1	N	50	34
2	n	10	10
3	$\overline{Y}$	555.4345	747.5882
4	$\overline{X}$	878.1624	208.8824
5	S_x	1084.678	150.506
6	S_y	584.826	443.9541
7	S_z	100.4271	446.8028
8	ρ_{yx}	0.8038341	0.9092764
9	ρ_{zx}	0.7922381	0.9087477

tors in comparison to RRT mean estimator is determined using the following equation

$$\text{PRE}(t_i, t_0) = \frac{\text{MSE}(t_0)}{\text{MSE}(t_i)} * 100, \quad \text{where} \quad i = 0, R, Reg, ER, S. \tag{2.5.1}$$

Table 2.2 consists of MSE and PRE of proposed and competing estimators for the real populations calculated using Equation (2.5.1). The following Figure 2.1 illustrates

Table 2.2: MSE and PRE of Estimators for both the real population

Estimators	MSE	PRE	MSE	PRE
t_0	27581.53	100.0000	13994.55	100.0000
t_R	14219.63	193.968	3710.477	377.1631
t_{Reg}	10488.81	262.9615	2424.338	577.2524
t_{ER}	11780.67	234.1253	3698.397	378.395
t_S	10280.7	268.2846	2415.272	579.4192

the PRE of the proposed estimator over the above existing estimators for both the real populations.

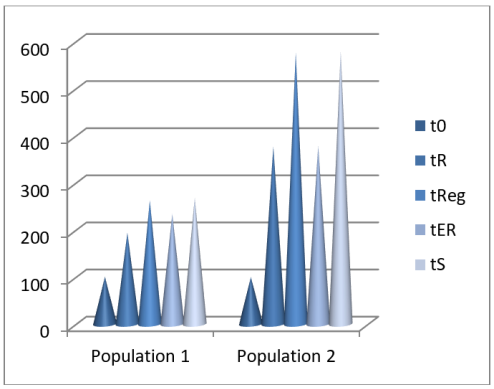

Figure 2.1: PRE of proposed over other estimators

2.6 Simulation Study

In the current section, a simulation inquiry is performed to evaluate the effectiveness of the proposed estimator. Two populations have been considered with distinct means and variance-covariance matrices to outline the distributions of the subsidiary variable and sensitive variable. The first data set is Artificial Population (AP1) generated from a bivariate normal distribution with mean and variance-covariance matrix

$$\mu = [3,3] \quad \text{and} \quad \Sigma = \begin{bmatrix} 12 & 3 \\ 3 & 6 \end{bmatrix}.$$

Another data set is Artificial Population (AP2) generated from a bivariate normal distribution that contains the parameters of the real population 2, considered in Section 2.5, with mean and covariance matrix

$$\mu = [747.5882, 208.8824] \quad \text{and} \quad \Sigma = \begin{bmatrix} 197095.3 & 60755.81 \\ 60755.81 & 22652.05 \end{bmatrix}.$$

We examined a population of fixed size 1000 taken from the above mentioned arti-

ficial populations AP1 and AP2. The scrambled variate S is deemed to follow a normal distribution that contains a mean equivalent to zero and a standard deviation equivalent to 10% of the standard deviation of auxiliary variable X. The response of the respondent is collected considering the additive model $Z = Y + S$. Different samples of sizes $n = 20,\ 50,\ 100,\ 200,\ 300$ are considered to ensure the efficiency of the estimator. Tables 2.3 and 2.4 contain the values of mean square error and percent relative efficiencies of t_R, t_{Reg}, t_{ER}, and ts with respect to t_0 for the populations AP1 and AP2, respectively.

The above table illustrates the performance of the suggested and previously known estimators for different sample sizes. Figure 2.2 and Figure 2.3 illustrate the maximum PRE of the suggested estimator over the competing estimators under consideration with a fixed sample size for Artificial Population 1 and Artificial Population 2.

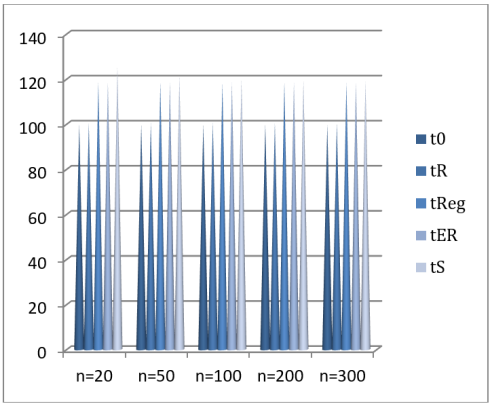

Figure 2.2: PRE of proposed over other estimators for Artificial Population 1

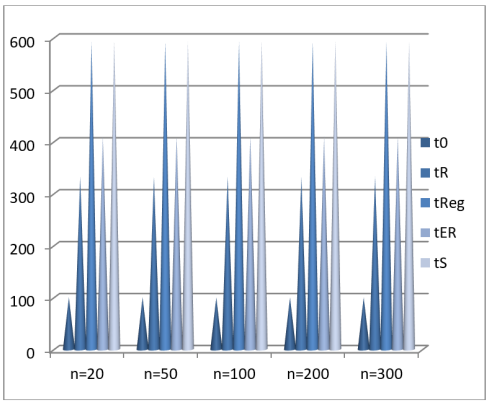

Figure 2.3: PRE of proposed over other estimators for Artificial Population 2

Table 2.3: MSE and PRE of Artificial Population 1

N=1000	Estimators	MSE	PRE
n=20	t_0	0.5608795	100.0000
	t_R	0.5578836	100.537
	t_{Reg}	0.4732741	118.5105
	t_{ER}	0.4732747	118.5103
	t_S	0.4477979	125.2528
n=50	t_0	0.2174839	100.0000
	t_R	0.2163222	100.537
	t_{Reg}	0.1835145	118.5105
	t_{ER}	0.1835147	118.5103
	t_S	0.1791823	121.3758
n=100	t_0	0.1030187	100.0000
	t_R	0.1024684	100.537
	t_{Reg}	0.0869279	118.5105
	t_{ER}	0.086928	118.5104
	t_S	0.08578929	120.0834
n=200	t_0	0.04578608	100.0000
	t_R	0.04554152	100.537
	t_{Reg}	0.03863462	118.5105
	t_{ER}	0.03863467	118.5103
	t_S	0.03833486	119.4372
n=300	t_0	0.02670855	100.0000
	t_R	0.02656589	100.537
	t_{Reg}	0.02253686	118.5105
	t_{ER}	0.02253689	118.5104
	t_S	0.0224024	119.2218

2.7 Results and Discussion

Table 2.1 presents the MSE and PRE of the proposed and previously known estimators in quantitative terms. It is evident that the MSE of the proposed estimator t_s are 10280.7 and

Table 2.4: MSE and PRE of Artificial Population 2

N=1000	**Estimators**	**MSE**	**PRE**
n=20	t_0	9455.474	100.0000
	t_R	2843.554	332.5231
	t_{Reg}	1597.437	591.9153
	t_{ER}	2306.867	409.8838
	t_S	1591.473	594.1335
n=50	t_0	3666.408	100.0000
	t_R	1102.603	332.5229
	t_{Reg}	619.4143	591.9153
	t_{ER}	894.4994	409.8838
	t_S	618.1991	593.0788
n=100	t_0	1736.72	100.0000
	t_R	522.2855	332.5231
	t_{Reg}	293.4068	591.9154
	t_{ER}	423.7102	409.8839
	t_S	293.0048	592.7275
n=200	t_0	771.8754	100.0000
	t_R	232.1269	332.523
	t_{Reg}	130.403	591.9154
	t_{ER}	188.3157	409.8837
	t_S	130.263	592.5515
n=300	t_0	450.2607	100.0000
	t_R	450.2607	332.523
	t_{Reg}	76.06843	591.9153
	t_{ER}	109.8508	409.8839
	t_S	75.99427	592.493

579.4192 for the real populations 1 and 2, respectively, which are smaller in comparison to the MSE of the previously described estimators of a population mean for both the real populations. The results are also supported by Figure 2.1, which demonstrates the PREs of the proposed estimator over the other competing estimators given in the literature. Tables

2.3 and 2.4 include quantitative measurements of MSE and PRE of the suggested and previously known estimators for different sample sizes for the simulated data AP1 and AP2, respectively. From the results given in Tables 2.3 and 2.4, we can infer that the MSE values for the proposed estimator are lesser than the MSE values for other estimators in the literature for given sample sizes. Figures 2.2 and 2.3 interpret the PRE of the presented estimator with the other listed estimators for the simulated populations AP1 and AP2, respectively.

2.8 Conclusion

From the above study, we acknowledged that being a biased estimator, ratio estimators perform better than RRT estimators since they have lower MSE than the ordinary RRT estimators, therefore have higher PRE in contrast with the usual RRT population mean estimator. The Bias and MSE of the proposed estimator are derived using the approximation of first order. At $k = 1$, the recommended estimator changes to Gupta *et al.* (2012). The proposed estimator has been compared with the RRT sample mean, ratio estimator, regression estimator and exponential ratio type estimator and the conditions for the recommended estimator to be better than the competing ones are discussed. The suggested estimator exhibits a percentage relative efficiency greater than 100, indicating higher efficiency than the RRT estimator. Two real and simulated data sets are used to support the theoretical results. Table 2.1 includes the parametric values of the real data sets. The findings from Tables 2.2, 2.3 and 2.4 reveal that the suggested estimator has the maximum PRE, hence it is the most efficient among the competing estimators under consideration.

Thus, the suggested estimator can be utilized to improve the population mean estimates using a known non-sensitive auxiliary variable. In addition, the recommended scrambled estimator may be used for a variety of sample strategies and can be employed in various sensitive and confidential situations in real life, including agricultural, medical, biology, economics, business and management areas.

Chapter 3

Generalized Classes of Estimators for Population Mean of Sensitive Variable Using Non-sensitive Auxiliary Parameters

3.1 Introduction

In this Chapter, we generalize the Sousa *et al.* (2010) family of estimators using some known auxiliary parameters. Another generalized family of estimators employing Searls (1964) methodology is proposed, which is considered to perform better than above mentioned family of estimators. The remaining parts of the paper are organized as follows- In Section 2, we examine various estimators of the finite population mean that are accessible in literature. Section 3 provides the bias and mean squared error up to the first order of approximation of the suggested class of estimators along with the optimum condition and minimum MSE. Another generalized family of estimators along with the corresponding

bias and MSE expressions are given in Section 4. The theoretical efficiency comparisons of the proposed estimators with competing estimators are presented in Section 5, and the efficiency criteria over competing estimators are derived. To examine the performances of the members of the suggested classes of estimators, an empirical study is carried out in Section 6. A simulation study is performed in Section 7 to showcase the effectiveness of various members of the suggested and competing classes of estimators. Section 8 presents the findings of the study as well as a discussion. Finally, some concluding remarks are given in Section 9.

3.2 Some Existing Estimators

Let Y be the research variable, which comprises delicate features that may not be accessed exactly as a result of the respondent's response. Let X be a non-sensitive secondary variable that is positively correlated to Y. Let S be a scrambled variable with a predefined distribution which is uncorrelated with Y and X. The sensitive variable Y is scrambled using the variable S. To assess Y, each respondent is instructed to choose a random number from the S distribution, say s, and add it to the actual value of Y. Let Z denote the reported scrambled response to Y that was initially suggested by Warner (1965) and later elaborated by Pollock and Bek (1976), given by

$$Z = Y + S.$$

Consider a simple random sample of size n drawn without replacement from the

population U to estimate the population mean of the sensitive research variable Y. Let y_i and x_i, $(i = 1,2,\ldots,n)$ represent the sample values of the research and non-sensitive supplementary variables on the i^{th} units, respectively. To obtain the bias and MSE of the proposed estimator, we define the following relative error terms,

$$e_0 = \frac{(\overline{z} - \overline{Z})}{\overline{Z}} \quad \text{and} \quad e_1 = \frac{(\overline{x} - \overline{X})}{\overline{X}},$$

such that $E(e_0) = E(e_1) = 0, \quad E(e_0^2) = \lambda C_z^2, \quad E(e_1^2) = \lambda C_x^2 \quad \text{and} \quad E(e_0 e_1) = \lambda \rho_{zx} C_z C_x,$

where $\lambda = (1-f)/n$ and $f = n/N$.

When the characteristic provided on secondary variable X is not taken into account, the typical RRT sample mean is an unbiased estimator of the population mean of a sensitive variable, given by

$$t_0 = \overline{z}. \tag{3.2.1}$$

The Variance of t_0 is given by

$$\mathrm{V}(t_0) = \lambda (S_y^2 + S_s^2),$$

where $\lambda = \dfrac{N-n}{Nn}$, $S_y^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2$ and $S_s^2 = \dfrac{1}{N-1}\sum_{i=1}^{N}(s_i - \overline{S})^2$.

Using the available non-sensitive auxiliary information, Sousa *et al.* (2010) improved the RRT estimator of the population mean of the sensitive variable by introducing ratio estimator given by

$$t_R = \overline{z}\left(\frac{\overline{X}}{\overline{x}}\right). \tag{3.2.2}$$

This estimator is biased for the population mean of the sensitive study variable. The bias and MSE of the aforementioned estimator, accurate up to the first degree of approximation are given by

$$\text{Bias}(t_R) = \lambda \overline{Z}(C_x^2 - \rho_{zx} C_z C_x), \text{ and}$$

$$\text{MSE}(t_R) = \lambda \overline{Z}^2 (C_z^2 + C_x^2 - 2\rho_{zx} C_z C_x).$$

Sousa *et al.* (2010) further modified the ratio estimator by making use of some known auxiliary parameters, given by

$$t_{TR} = \overline{z} \left(\frac{c\overline{X} + d}{c\overline{x} + d} \right), \tag{3.2.3}$$

where c and d are auxiliary parameters such as the coefficient of skewness, coefficient of kurtosis, etc. The bias and MSE of the modified ratio estimator are given by

$$\text{Bias}(t_{TR}) = \lambda \overline{Z}(\phi^2 C_x^2 - \phi \rho_{zx} C_z C_x), \text{ and}$$

$$\text{MSE}(t_{TR}) = \lambda \overline{Z}^2 (C_z^2 + \phi^2 C_x^2 - 2\phi \rho_{zx} C_z C_x).$$

where $\phi = \frac{c\overline{X}}{c\overline{X}+d}$.

Gupta *et al.* (2012) provided the typical regression estimator for the finite population mean of the susceptible study variable Y, as follows

$$t_{Reg} = \overline{z} + b_{zx}(\overline{X} - \overline{x}), \tag{3.2.4}$$

where $b_{zx} = \frac{s_{zx}}{s_x^2}$ is the sample regression coefficient of Z on X. Up to the first order of approximation, the bias of this regression estimator is given as

$$\text{Bias}(t_{Reg}) = -\lambda \beta_{zx} \left(\frac{\mu_{12}}{\mu_{11}} - \frac{\mu_{03}}{\mu_{02}} \right),$$

where $\beta_{zx} = \frac{S_{zx}}{S_x^2} = \frac{S_{yx}}{S_x^2} = \rho_{yx}\frac{S_y}{S_x} = \beta_{yx}$ is the population regression coefficient and $u_{rs} = E(z_i - \overline{Z})^r (x_i - \overline{X})^s$. The mean square error of the regression estimator, up to the first order of approximation is given as

$$\begin{aligned} \text{MSE}(t_{Reg}) =& \lambda \overline{Z}^2 C_Z^2 (1 - \rho_{zx}^2) \\ =& \lambda S_y^2 \left[\left(1 + \frac{S_s^2}{S_y^2} \right) - \rho_{zx}^2 \right]. \end{aligned}$$

3.3 Generalized class of ratio estimators

Several authors have developed more precise estimators by using certain known auxiliary parameters. Motivated from Khoshnevisan *et al.* 2007, we generalized ratio type estimators given by Sousa *et al.* (2010) under the SRS scheme as

$$t_r = \overline{z} \left[\frac{a\overline{X} + b}{\alpha(a\overline{x} + b) + (1 - \alpha)(a\overline{X} + b)} \right]^g, \tag{3.3.1}$$

where $a(\neq 0)$ and b are either real numbers or functions of the known parameters of the non-sensitive auxiliary variable X such as skewness $\beta_1(x)$, kurtosis $\beta_2(x)$, standard deviation σ, coefficient of variation C_x and correlation coefficient ρ.

To estimate the bias and mean square error of t_r, we consider the following assumption

$$\bar{z} = \overline{Z}(1+e_0) \quad \text{and} \quad \bar{x} = \overline{X}(1+e_1),$$

such that $E(e_0) = E(e_1) = 0$, $E(e_0^2) = \lambda C_z^2$, $E(e_1^2) = \lambda C_x^2$, $E(e_0 e_1) = \lambda \rho C_z C_x$, where $C_z^2 = \dfrac{S_z^2}{\overline{Z}^2}$ and $C_x^2 = \dfrac{S_x^2}{\overline{X}^2}$.

Expressing (3.3.1) in terms of $e_i's$, we have

$$\begin{aligned}
t_r &= \overline{Z}(1+e_0)\left[\frac{a\overline{X}+b}{\alpha(a\overline{X}(1+e_1)+b)+(1-\alpha)(a\overline{X}+b)}\right]^g \\
&= \overline{Z}(1+e_0)\left[\frac{a\overline{X}+b}{\alpha(a\overline{X}+a\overline{X}e_1+b)+(1-\alpha)(a\overline{X}+b)}\right]^g \\
&= \overline{Z}(1+e_0)\left[\frac{a\overline{X}+b}{(a\overline{X}+b)\left(\alpha\left(1+\frac{a\overline{X}}{a\overline{X}+b}e_1\right)+(1-\alpha)\right)}\right]^g \\
&= \overline{Z}(1+e_0)\left[\frac{1}{(\alpha(1+\phi e_1)+(1-\alpha))}\right]^g \\
&= \overline{Z}(1+e_0)(1+\alpha\phi e_1)^{-g}, \qquad (3.3.2)
\end{aligned}$$

where $\phi = \dfrac{a\overline{X}}{a\overline{X}+b}$. Assuming $|\alpha\phi e_1| < 1$ so that $(1+\alpha\phi e_1)^{-g}$ is expandable, we expand the right hand side of (3.3.2) and consider terms up to the first order of approximation, given by

$$\begin{aligned}
t_r &= \overline{Z}\left[1+e_0-\alpha\phi g e_1+\frac{g(g+1)}{2}\alpha^2\phi^2 e_1^2-\alpha\phi g e_0 e_1\right] \\
t_r-\overline{Z} &= \overline{Z}\left[e_0-\alpha\phi g e_1+\frac{g(g+1)}{2}\alpha^2\phi^2 e_1^2-\alpha\phi g e_0 e_1\right]. \qquad (3.3.3)
\end{aligned}$$

Taking expectations on both sides of the above Equation (3.3.3), we get the bias of

the estimator up to the first order of approximation, given by

$$\text{Bias}(t_r) = \lambda \overline{Z} \left[\frac{g(g+1)}{2} \alpha^2 \phi^2 C_x^2 - \alpha \phi g \rho_{zx} C_z C_x \right]. \tag{3.3.4}$$

Squaring both the side of the (3.3.3) and then taking expectations on both sides, we get MSE of the proposed class of estimators up to the first order of approximation, as

$$\begin{aligned} \text{MSE}(t_r) &= E(t_r - \overline{Z})^2 \\ &= \overline{Z}^2 \left[e_0 - \alpha \phi g e_1 + \frac{g(g+1)}{2} \alpha^2 \phi^2 e_1^2 - \alpha \phi g e_0 e_1 \right]^2 \\ &= \overline{Z}^2 \left[e_0^2 - \alpha^2 \phi^2 g^2 e_1^2 - 2\alpha \phi g e_0 e_1 \right] \\ &= \lambda \overline{Z}^2 \left[C_z^2 + \alpha^2 \phi^2 g^2 C_x^2 - 2\alpha \phi g \rho_{zx} C_z C_x \right]. \end{aligned} \tag{3.3.5}$$

Minimizing MSE(t_r)w.r.t α, we obtain the optimum value of α as

$$\alpha = \frac{\rho_{zx} C_z}{\phi g C_x}.$$

Substituting the optimal value of α in (3.3.5), we get the minimum value of MSE, given as

$$\text{MSE}_{\min}(t_r) = \lambda \overline{Z}^2 C_z^2 (1 - \rho_{zx}^2). \tag{3.3.6}$$

The minimum value of MSE(t_r) is the same as the estimated variance of the typical linear regression estimator for finite population mean of the sensitive study variable. For the generalized class of estimators given in Table 3.1, we can express the MSE given in

(3.3.5) by the following equation

$$\text{MSE}(t_i) = \begin{cases} \overline{Z}^2 \lambda \left(C_y^2 + C_s^2\right), & \text{for } i = 0, \\ \overline{Z}^2 \lambda \left(C_z^2 + C_x^2 - 2\rho_{zx} C_z C_x\right), & \text{for } i = 1, \\ \overline{Z}^2 \lambda \left(C_z^2 + \phi_i^2 C_x^2 - 2\phi_i \rho_{zx} C_z C_x\right), & \text{for } i = 2, 3, \ldots, 12. \end{cases}$$

Table 3.1: Members of t_r family of estimators

S. No.	Estimators	α	a	b	g
1	$t_o = \overline{z}$	0	0	0	0
2	$t_1 = \overline{z}\left(\frac{\overline{X}}{\overline{x}}\right)$	1	1	0	1
3	$t_2 = \overline{z}\left(\frac{\overline{X}+C_x}{\overline{x}+C_x}\right)$	1	1	C_x	1
4	$t_3 = \overline{z}\left(\frac{\beta_2\overline{X}+C_x}{\beta_2\overline{x}+C_x}\right)$	1	β_2	C_x	1
5	$t_4 = \overline{z}\left(\frac{\overline{X}+\beta_1}{\overline{x}+\beta_1}\right)$	1	1	β_1	1
6	$t_5 = \overline{z}\left(\frac{C_x\overline{X}+\beta_2}{C_x\overline{x}+\beta_2}\right)$	1	C_x	β_2	1
7	$t_6 = \overline{z}\left(\frac{\overline{X}+S_x}{\overline{x}+S_x}\right)$	1	1	S_x	1
8	$t_7 = \overline{z}\left(\frac{\beta_1\overline{X}+S_x}{\beta_1\overline{x}+S_x}\right)$	1	β_1	S_x	1
9	$t_8 = \overline{z}\left(\frac{\beta_2\overline{X}+S_x}{\beta_2\overline{x}+S_x}\right)$	1	β_2	S_x	1
10	$t_9 = \overline{z}\left(\frac{\overline{X}+\rho}{\overline{x}+\rho}\right)$	1	1	ρ	1
11	$t_{10} = \overline{z}\left(\frac{\overline{X}+\beta_2}{\overline{x}+\beta_2}\right)$	1	1	β_2	1
12	$t_{11} = \overline{z}\left(\frac{\beta_1\overline{X}+\beta_2}{\beta_1\overline{x}+\beta_2}\right)$	1	β_1	β_2	1
13	$t_{12} = \overline{z}\left(\frac{\beta_2\overline{X}+\beta_1}{\beta_2\overline{x}+\beta_1}\right)$	1	β_2	β_1	1

3.4 Improved generalized class of estimators

Considering the technique given in Searls 1964, which produced a biased sample mean estimator with lower MSE than the usual sample mean $\overline{y}$, we suggest a novel searls type estimator to estimate the population mean of the sensitive study variable which performs

better than the above existing estimators in the literature as,

$$z_r = k_r \bar{z} \left[\frac{a\overline{X} + b}{\alpha(a\bar{x} + b) + (1-\alpha)(a\overline{X} + b)} \right]^g, \tag{3.4.1}$$

where k_r is an appropriate constant to be found in order to minimize the MSE of the z_r estimator. Using the Taylor series expansion and expressing z_r in terms of $e_i's$, we get

$$\begin{aligned}
z_r &= k_r \overline{Z}(1+e_0) \left[\frac{a\overline{X} + b}{\alpha(a\overline{X}(1+e_1) + b) + (1-\alpha)(a\overline{X} + b)} \right]^g \\
&= k_r \overline{Z}(1+e_0) \left[\frac{a\overline{X} + b}{\alpha(a\overline{X} + a\overline{X}e_1 + b) + (1-\alpha)(a\overline{X} + b)} \right]^g \\
&= k_r \overline{Z}(1+e_0) \left[\frac{a\overline{X} + b}{(a\overline{X} + b)\left(\alpha\left(1 + \frac{a\overline{X}}{a\overline{X}+b}e_1\right) + (1-\alpha)\right)} \right]^g \\
&= k_r \overline{Z}(1+e_0) \left[\frac{1}{(\alpha(1+\phi e_1) + (1-\alpha))} \right]^g \\
&= k_r \overline{Z}(1+e_0)(1+\alpha\phi e_1)^{-g}.
\end{aligned} \tag{3.4.2}$$

Preserving its terms up to the first degree of approximation and substracting $\overline{Z}$ from both the sides of (3.4.2), we get

$$z_r - \overline{Z} = k_r \overline{Z} \left[1 - g\alpha\phi e_1 + \frac{g(g+1)}{2}\alpha^2\phi^2 e_1^2 + e_0 - g\alpha\phi e_0 e_1 \right] - \overline{Z}. \tag{3.4.3}$$

Taking the expectations on both sides of (3.4.3), we get the bias of the estimator z_r as

$$\text{Bias}(z_r) = k_r \overline{Z}\lambda \left[\frac{g(g+1)}{2}\alpha^2\phi^2 C_x^2 - g\alpha\phi C_{zx} \right] + \overline{Z}(k_r - 1). \tag{3.4.4}$$

Squaring both the sides and then taking expectations on both sides of the Equation (3.4.3), we get the MSE of z_r as

$$\text{MSE}(z_r) = E(z_r - \overline{Z})^2$$

$$
\begin{aligned}
&= E\left[k_r\overline{Z}\left\{1-g\alpha\phi e_1+\frac{g(g+1)}{2}\alpha^2\phi^2e_1^2+e_0-g\alpha\phi e_0e_1\right\}-\overline{Z}\right]^2.\\
&= E\left[k_r^2\overline{Z}^2\left\{1-g\alpha\phi e_1+\frac{g(g+1)}{2}\alpha^2\phi^2e_1^2+e_0-g\alpha\phi e_0e_1\right\}^2+\overline{Z}^2\right.\\
&\quad\left.-2k_r\overline{Z}^2\left\{1-g\alpha\phi e_1+\frac{g(g+1)}{2}\alpha^2\phi^2e_1^2+e_0-g\alpha\phi e_0e_1\right\}\right]\\
&= \overline{Z}^2\left[(k_r-1)^2+k_r^2\lambda C_z^2+(k_r^2(2g^2+g)-k_r(g^2+g))\alpha^2\phi^2\lambda C_x^2\right.\\
&\quad\left.-2g\alpha\phi(2k_r^2-k_r)C_{zx}\right]. \qquad (3.4.5)
\end{aligned}
$$

On minimizing $\mathrm{MSE}(z_r)$, we obtain the optimum value of k_r as

$$k_r=\frac{P}{2Q}, \qquad (3.4.6)$$

where

$$P=2+g(g+1)\alpha^2\phi^2\lambda C_x^2-2g\alpha\phi\lambda C_{zx}, \quad \text{and}$$

$$Q=1+\lambda C_z^2+2g(g+1)\alpha^2\phi^2\lambda C_x^2-4g\alpha\phi\lambda C_{zx}.$$

After substituting the optimum value of k_r in (3.4.5), the minimum MSE of z_r is given by

$$\mathrm{MSE}_{\min}(z_r)=\overline{Z}^2\left[1-\frac{P^2}{4Q}\right]. \qquad (3.4.7)$$

The Mean Square Error given in (3.4.5) can be expressed for the ratio estimators in Table 3.2 by the following equation

$$
\mathrm{MSE}(z_i)=\begin{cases}
\overline{Z}^2\left(k_i^2\lambda C_z^2+(k_i-1)^2\right), & i=0,\\
\overline{Z}^2\left(k_i^2\lambda C_z^2+(3k_i^2-2k_i)C_x^2-2(2k_i^2-k_i)\lambda C_{zx}+(k_i-1)^2\right), & i=1,\\
\overline{Z}^2\left(k_i^2\lambda C_z^2+(3k_i^2-2k_i)\phi_i^2\lambda C_x^2-2(2k_i^2-k_i)\phi_i\lambda C_{zx}+(k_i-1)^2\right), & i=2,3,\ldots,12,
\end{cases} \qquad (3.4.8)
$$

The optimal value of k_i, for which the $\mathrm{MSE}(z_i)$ is minimized, is given by

Table 3.2: Members of z_r family of estimators

S. No.	Estimators	α	a	b	g
1	$z_o = k_0\overline{z}$	0	0	0	0
2	$z_1 = k_1\overline{z}\left(\frac{\overline{X}}{\overline{x}}\right)$	1	1	0	1
3	$z_2 = k_2\overline{z}\left(\frac{\overline{X}+C_x}{\overline{x}+C_x}\right)$	1	1	C_x	1
4	$z_3 = k_3\overline{z}\left(\frac{\beta_2\overline{X}+C_x}{\beta_2\overline{x}+C_x}\right)$	1	β_2	C_x	1
5	$z_4 = k_4\overline{z}\left(\frac{\overline{X}+\beta_1}{\overline{x}+\beta_1}\right)$	1	1	β_1	1
6	$z_5 = k_5\overline{z}\left(\frac{C_x\overline{X}+\beta_2}{C_x\overline{x}+\beta_2}\right)$	1	C_x	β_2	1
7	$z_6 = k_6\overline{z}\left(\frac{\overline{X}+S_x}{\overline{x}+S_x}\right)$	1	1	S_x	1
8	$z_7 = k_7\overline{z}\left(\frac{\beta_1\overline{X}+S_x}{\beta_1\overline{x}+S_x}\right)$	1	β_1	S_x	1
9	$z_8 = k_8\overline{z}\left(\frac{\beta_2\overline{X}+S_x}{\beta_2\overline{x}+S_x}\right)$	1	β_2	S_x	1
10	$z_9 = k_9\overline{z}\left(\frac{\overline{X}+\rho}{\overline{x}+\rho}\right)$	1	1	ρ	1
11	$z_{10} = k_{10}\overline{z}\left(\frac{\overline{X}+\beta_2}{\overline{x}+\beta_2}\right)$	1	1	β_2	1
12	$z_{11} = k_{11}\overline{z}\left(\frac{\beta_1\overline{X}+\beta_2}{\beta_1\overline{x}+\beta_2}\right)$	1	β_1	β_2	1
13	$z_{12} = k_{12}\overline{z}\left(\frac{\beta_2\overline{X}+\beta_1}{\beta_2\overline{x}+\beta_1}\right)$	1	β_2	β_1	1

$$k_i = \begin{cases} \frac{1}{1+\lambda C_z^2}, & \text{for } i = 0, \\ \frac{1+\lambda C_x^2-\lambda C_{zx}}{1+3\lambda C_x^2-4\lambda C_{zx}+\lambda C_z^2}, & \text{for } i = 1, \\ \frac{1+\lambda\phi_i^2 C_x^2-\lambda\phi_i C_{zx}}{1+3\lambda\phi_i^2 C_x^2-4\lambda\phi_i C_{zx}+\lambda C_z^2}, & \text{for } i = 2,3,\ldots,12. \end{cases}$$

Putting the optimal value of k_i in (3.4.8) to obtain minimum value of MSE (z_i), we get

$$\text{MSE}_{\min}(z_i) = \begin{cases} \overline{Z}^2\left(1-\frac{1}{(1+\lambda C_z^2)}\right), & \text{for } i = 0, \\ \overline{Z}^2\left(1-\frac{A^2}{B}\right), & \text{for } i = 1, \\ \overline{Z}^2\left(1-\frac{C^2}{D}\right), & \text{for } i = 2,3,\ldots,12, \end{cases}$$

where

$$A = 1+\lambda C_x^2-\lambda C_{zx},$$

$$B = 1 + 3\lambda C_x^2 - 4\lambda C_{zx} + \lambda C_z^2,$$

$$C = 1 + \lambda \phi_i^2 C_x^2 - \lambda \phi_i C_{zx},$$

$$D = 1 + 3\lambda \phi_i^2 C_x^2 - 4\lambda \phi_i C_{zx} + \lambda C_z^2.$$

3.5 Efficiency comparisons

Under this section, we conceptually compare the proposed estimators against competing estimators of the population mean of the sensitive study variable using the known auxiliary parameters of the non-sensitive supplementary variable. The efficiency conditions of the introduced estimators over competing estimators under which the proposed estimators outperform competing estimators are obtained.

1. $\text{MSE}(t_1) < \text{MSE}(t_0)$, if

$$\overline{Z}^2 \lambda \left(C_z^2 + C_x^2 - 2\rho_{zx} C_z C_x \right) - \lambda C_z^2 \geq 0.$$

2. $\text{MSE}(t_i) < \text{MSE}(t_1)$, $i = 2, \ldots, 12$, if

$$\rho_{zx} > \frac{\phi_i C_x}{2C_z}, \quad \text{for} \quad \phi_i > 1,$$

$$\rho_{zx} < \frac{\phi_i C_x}{2C_z}, \quad \text{for} \quad \phi_i < 1.$$

3. $\text{MSE}(z_0) < \text{MSE}(t_0)$, if

$$\lambda C_z^2 - \left[1 - \frac{1}{(1 + \lambda C_z^2)} \right] > 0.$$

4. $\text{MSE}(z_1) < \text{MSE}(t_1)$, if

$$\lambda\left(C_z^2 + C_x^2 - 2\rho_{zx}C_zC_x\right) - \left(1 - \frac{A^2}{B}\right) > 0.$$

5. $\text{MSE}(z_i) < \text{MSE}(t_i)$, $i = 2, \ldots, 12$, if

$$\lambda\left(C_z^2 + \phi_i^2 C_x^2 - 2\phi_i\rho_{zx}C_zC_x\right) - \left(1 - \frac{C^2}{D}\right) > 0.$$

6. $\text{MSE}(z_r) < \text{MSE}(t_r)$, if

$$\lambda C_z^2(1 - \rho_{zx}^2) - \left(1 - \frac{P^2}{4Q}\right) > 0.$$

3.6 Numerical Illustration

Two real populations are incorporated into the study to validate the theoretical results. Population 1 is taken from Sarndal and Wretman 1986, where X denotes the population during 1985 and Y denotes the revenue from the 1985 municipal taxation. Population 2 is taken from Murthy 1967, where X denotes the number of workers in the factory and Y denotes the output for 80 factories in a region. The parametric specifications of both the populations under consideration are provided in Table 3.3 given below.

With a view to analyzing the competency of the proposed families of estimators, we have assessed the percent relative efficiencies of all the competing estimators under consideration for both populations. PRE for two proposed families of estimators in comparison

Table 3.3: Descriptive Statistics of the Real Population

S. No.	Information	Population 1	Population 2
1	N	250	50
2	n	15	10
3	$\overline{Y}$	30.944	555.4345
4	$\overline{X}$	259.276	878.1624
5	S_x	630.8737	1084.678
6	S_y	54.28762	584.826
7	S_z	630.8737	587.1704
8	ρ_{zx}	0.6476499	0.7922381

to RRT usual mean estimator is determined using the following equation:

$$\text{PRE}(\cdot, \mu_0) = \frac{\text{MSE}(\mu_0)}{\text{MSE}(\cdot)} \times 100. \tag{3.6.1}$$

Table 3.4 demonstrates the effectiveness of the proposed generalized classes of estimators over the generalized ratio estimator for both the real populations and infer that for Population 1, z_1 is the most efficient among all the estimators and for population 2, z_8 is the most efficient among the class of all competing estimators. Figures 3.1 and 3.2 demonstrate the PRE values of the recommended classes over the t_0 estimator for the real data sets 1 and 2, respectively.

Table 3.4: MSE and PRE of t_r and z_r families of estimators

S. No.	MSE	PRE	MSE	PRE
t_0	421.8243	100.0000	27581.53	100.0000
t_1	269.3494	156.6086	14219.63	193.968
t_2	258.0977	163.4359	14377.67	191.8359
t_3	260.0048	162.2371	14403.72	191.4889
t_4	253.8027	166.2017	14368.81	191.9542
t_5	240.7016	175.2478	14314.16	192.6871
t_6	282.7648	149.1785	12788.8	215.6693
t_7	229.8445	183.526	11244.25	245.2946
t_8	254.0706	166.0264	10988.6	251.0014
t_9	259.5061	162.5489	14389.71	191.6753
t_{10}	229.3961	183.8847	14291.6	192.9912
t_{11}	252.9566	166.7576	14334.31	192.4162
t_{12}	259.9447	162.2746	14401.71	191.5156
t_{Reg}	227.5791	185.3528	10630.23	259.4632
z_0	545.9258	241.667	25978.26	106.8563
z_1	164.2812	**256.7697**	12422.74	222.0245
z_2	164.8212	255.9284	12403.72	222.3649
z_3	164.2881	256.7589	12418.44	222.1014
z_4	166.061	254.0177	12398.71	222.4548
z_5	170.4931	247.4143	12367.78	223.0111
z_6	230.5033	183.0014	12548.73	219.7953
z_7	177.0566	238.2426	11015.2	250.3953
z_8	165.9818	254.1389	10403.72	**265.1122**
z_9	164.4266	256.5426	12410.53	222.243
z_{10}	177.5914	237.5252	12355	223.2418
z_{11}	166.3131	253.6326	12379.2	222.8055
z_{12}	164.3048	256.7329	12417.3	222.1217

3.7 Simulation Study

Under this section, we analyze the performance of different estimators fallen under both the classes of estimators mentioned in Tables 3.1 and 3.2. We compared all the estimators with the usual unbiased RRT estimator with a prototypical simulation study. Population 1 and Population 2 are artificial populations generated from bivariate normal distributions

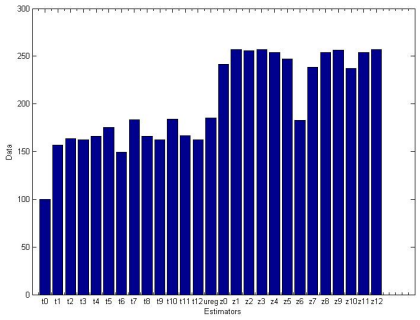

Figure 3.1: PRE values of recommended classes of estimators for real data set 1

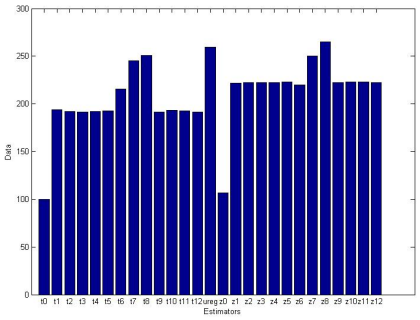

Figure 3.2: PRE values of recommended classes of estimators for real data set 2

with presumed mean vectors

$$\mu_1 = [3,3] \qquad \text{and} \quad \mu_2 = [2,2], \quad \text{respectively,}$$

and variance-covariance matrices

$$\Sigma_1 = \begin{bmatrix} 50 & 9 \\ 9 & 16 \end{bmatrix} \quad \text{and} \quad \Sigma_2 = \begin{bmatrix} 64 & 9 \\ 9 & 5 \end{bmatrix}, \quad \text{respectively.}$$

The comparison of the proposed families of estimators is made with respect to the traditional RRT mean estimator. The MSE and PRE values of the estimators obtained using Equation (3.6.1) are given in Table 3.5. Figures 3.3 and 3.4 illustrate the PRE values of the recommended classes over the t_0 estimator.

Table 3.5: MSE and PRE of t_r and z_r families of estimators

S. No.	MSE	PRE	MSE	PRE
t_0	0.4356	100.0000	1.1909	100.0000
t_1	0.4005	108.7686	0.9101	130.8436
t_2	0.3723	117.0196	0.9906	120.2167
t_3	0.3817	114.1146	0.9412	126.5241
t_4	0.3981	109.4217	0.9055	131.5139
t_5	0.3718	117.173	1.03474	115.0918
t_6	0.3782	115.1699	1.0338	115.1952
t_7	0.4317	100.9124	1.1839	100.5864
t_8	0.3724	116.9909	0.9672	123.1255
t_9	0.3850	113.158	0.9503	125.3144
t_{10}	0.3737	116.5844	1.0493	113.4866
t_{11}	0.431616	100.9433	1.1886	100.5926
t_{12}	0.3982	109.4031	0.9056	131.5033
t_{Reg}	0.37181	117.1789	0.8279	143.8297
z_0	0.4135	105.3598	0.8907	133.7031
z_1	106.9926	116.3743	0.74802	159.0434
z_2	0.3515	**122.3090**	0.7996	148.9237
z_3	0.36244	120.2067	0.7714	154.3629
z_4	0.3743	116.3978	0.7487	**159.0593**
z_5	0.3562	122.2666	0.8223	144.8089
z_6	0.3631	119.9391	0.8219	144.8895
z_7	0.4109	106.031	0.8864	134.3508
z_8	0.35625	122.2962	0.7867	151.3747
z_9	0.36476	119.4448	0.7769	153.2824
z_{10}	0.3586	121.4686	0.8295014	143.5683
z_{11}	0.41079	106.0603	0.8863849	134.3549
z_{12}	0.3743	116.3825	0.7487	159.0489

3.8 Results and Discussion

Tables 3.1 and 3.2 consist of the members of t_r and z_r families of estimators for different values of auxiliary parameters. From Table 3.4, it is observed that the proposed estimator z_1 and z_8 have the least values of MSE which are 164.2812 and 10403.72 for the real data sets 1 and 2 respectively. From Table 3.5, it is contemplated that the proposed estimator z_2

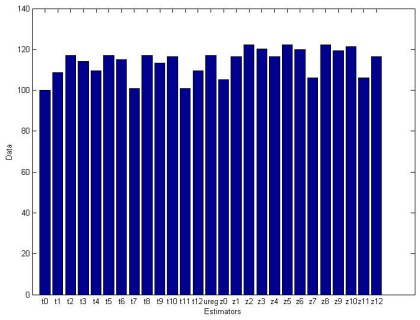

Figure 3.3: PRE values of recommended classes of estimators for simulated data set 1

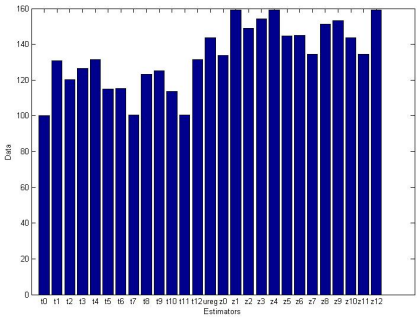

Figure 3.4: PRE values of recommended classes of estimators for simulated data set 2

and z_4 have the least values of MSE which are 0.3562185 and 0.7487154 for the artificial populations 1 and 2, respectively. From the above results, we can infer that both the proposed generalized families of estimators are more efficient than the traditional RRT mean estimator. The study also concluded that $\text{MSE}(z_i) < \text{MSE}(t_i)$, for all $i = 0, 1, 2, \ldots, 12$. From this, we also interpret that the proposed family of estimators given in Section 3.4 are more efficient than the generalized ratio family of estimators given in Section 3.3.

3.9 Conclusion

In this study, two generalized classes of estimators are introduced to estimate the population mean of a sensitive study variable using non-sensitive supplementary information under the simple random sampling design. According to the findings of the preceding study, we observe that being a biased estimator, ratio estimators outperform RRT mean estimator since they have a lower MSE, and thus have a higher PRE than the ordinary RRT mean estimator of the population mean of the sensitive variable. The performance of

the two proposed classes of estimators are compared to that of several existing estimators using two real life data sets. In addition, a simulation research employing two artificial data sets is carried out to examine the efficiencies and consistencies of the proposed families of estimators. From the results given in Tables 3.4 and 3.5, we can infer that the proposed class of estimators z_r is more efficient than the traditional ratio and regression estimators given by Sousa *et al.* (2010) and Gupta *et al.* (2012). Figures 3.1, 3.2, 3.3 and 3.4 indicate the PRE of recommended classes of estimators over t_0 estimator, for both real and simulated data sets. As a result, the proposed families of estimators can be utilized to improve the estimation of the population mean of a sensitive variable using a known non-sensitive auxiliary information in various sensitive areas of research in the Health and Social Sciences. Furthermore, the proposed family could be extended to alternative sampling strategies.

Chapter 4

Robust Type Regression Estimator of the Mean of a Sensitive Variable

4.1 Introduction

Since 1960, substantial theoretical work has been undertaken for developing statistical processes that are resistive to minor deviations from assumptions, i.e. robust towards outliers and stable to small variations from the presumed parametric model. In reality, it is widely known that classical optimal techniques perform fairly poorly when the rigorous model assumptions are violated. Furthermore, the outliers present in the data impose a negative impact on the efficiency of the traditional estimators as the Ordinary Least Square (OLS) estimators are susceptible to extreme values. Another approach is to discard the outliers and still use OLS method but this would not be efficient if the numbers of outliers are large. To remedy this problem, many authors have used alternative methods of regression. One such method is to use robust techniques as they are extremely helpful in detecting outliers in the data. Kadilar *et al.* (2007) incorporated robust technique and

replaced ordinary least square estimation with Huber M estimators in order to decrease the negative effects of outlier problem in the data. Oral and Kadilar (2011) incorporated modified maximum liklihood estimators into Kadilar *et al.* (2007) estimator and studied the robust properties of the modified estimators.

In addition to this, Zaman and Bulut (2019) modified ratio type estimators given by Sisodia and Dwivedi (1981) by adopting other robust techniques like Least Trimmed Square Method, Least Median of Square, Least Absolute Deviation, Tukey M, Hampel M, Huber M and Huber MM. Zaman (2019) improved upon the traditional methods given by Kadilar *et al.* (2007). Ali *et al.* (2021) generalized the work done by Zaman and Bulut (2019) for the sensitive study variable and presented a new regression type estimator using robust techniques. Subzar and Alanzi (2020) proposed novel ratio type estimators of finite population mean using simple random sampling without replacement incorporating the supplemental information in Bowley's coefficient of skewness. By substituting the quantile regression coefficient for the Ordinary least square regression coefficient, Anas *et al.* (2021) proposed an unique class of estimators by incorporating L-moment characteristics into existing estimators. Further by employing Searls (1964) technique, Grover and Kaur (2021) suggested an enhanced estimator of the population mean in simple random sampling with no replacement.

In this Chapter, we proposed a new robust regression type estimator using robust technique to improve the estimation of the population mean of a sensitive variable under simple random sampling scheme. The remains of the chapter are as follows- Section 2 comprises new methods of regression along with their properties. The existing competing

estimators have been discussed in Section 3. In Section 4, a new robust regression estimator has been proposed. The mean squared error of the introduced estimator are derived using Taylor series of expansion. The optimum value of the characterizing scalar is obtained by the Lagrange method of Maxima-Minima. The least value of the MSE of the suggested estimator is also obtained for this optimum value of the charactering constant. Theoretical results are verified by comparing the suggested robust estimator with Ali *et al.* (2021) using two real populations in Section 5. To assess the efficiency of the proposed estimator, a simulation study is performed in Section 6. The findings of the study and discussion are given in Section 7. Lastly, Section 8 provides the conclusion of the chapter. Let Y be the variate under study which is not precisely observable and let X be the supplementary variable which is directly related to the study variable Y. The scrambled variable S is independent of both Y and X variables. The records are collected using the additive model $Z = Y + S$ where respondents provide true information on X and scrambled information on Y. The scrambled variate S is assumed to follow normal distribution with mean zero and standard deviation equals to 10% of the standard deviation of X variable.

4.2 Alternative Methods of Regression

Alternative methods of regression are used when the traditional estimator fails to give reliable results. Such a situation occurs when data is contaminated and contain outliers that affect the efficiency of estimators. In the current section, different robust estimators and their properties are discussed.

4.2.1 M Estimator

The most commonly used robust method of estimation is M estimation suggested by Huber (1973). The class of M estimators are considered as a generalization of maximum likelihood estimators.

(i) **Huber M (hbm)**

The functional form of Huber M estimator is given as

$$\rho(y) = \begin{cases} \frac{1}{2}e^2, & \text{if} \quad |e| < k, \\ k|e| - \frac{1}{2}k^2, & \text{if} \quad |e| \geq k, \end{cases}$$

where k is the tuning constant equal to 1.345.

(ii) **Huber MM (hmm)**

MM method of estimation is a special case of M estimation. The method, introduced by Yohai (1987), has high breakdown point. Outliers are estimated as $e_i(T_0) = y_i - T_0' x_i, 1 \leq i \leq n$, where T_0 is taken as a starting point of estimation. Considering the constraints $b/a = 0.5$, where b is calculated using the following equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{e_i(\beta)}{S_n}\right) = b,$$

where S_n is M scale estimation and using ρ which meets the assumptions of Yohai (1987), α is regarded as $\max(\rho_0)$.

(iii) **Hampel M (hpm)**

This method, under M estimation, is given by Hampel (1971) with the functional form given by

$$\rho(y) = \begin{cases} \frac{y^2}{2}, & \text{if} \quad 0 < |y| < a, \\ a|y| - \frac{y^2}{2}, & \text{if} \quad a < |e| \leq b, \\ \frac{-a}{2(c-b)}(c-y)^2, & \text{if} \quad b < |e| \leq c, \\ \frac{a}{2}(b+c-a), & \text{if} \quad c < |y|, \end{cases}$$

where $a = 1.7, b = 3.4$, and $c = 8.5$.

(iv) **Tukey (tkm)**

The method was first suggested by Tukey (1977) and the functional form is given by

$$\rho(y) = \begin{cases} \frac{1}{6}\left(1 - \left(1 - \left(\frac{y}{k}\right)^2\right)^3\right), & \text{if} \quad |y| \leq k, \\ \frac{1}{6}, & \text{if} \quad |y| > k, \end{cases}$$

where k is taken as 5 or 6.

4.2.2 S Estimator

S method of estimation is a generalized method based on LMS and LTS methods of estimation. S estimators are considered to be more robust than the entire class of M estimators as they have smaller asymptotic bias. The functional form of S estimator is defined by

$$\min_\beta \sigma_s(e_1, e_2, ..., e_n),$$

where $\hat{\sigma}_s = \sqrt{\frac{1}{nK}\sum_{i=1}^{n} \omega_i e_i^2}$; $K = 0.199$ and $\omega_i = \frac{\rho(u_i)}{u_i^2}$.

4.2.3 Least Median of Square (lms)

The procedure takes into account the median of square of errors in place of sum of square of errors. The method was first suggested by Rousseeuw (1984). The function to be minimized is given as follows

$$\text{median}(\varepsilon_i^2).$$

The method is effective in the presence of the outliers in both X and Y directions with breakdown point of 0.5 or 50 (Rousseeuw and Leroy (1987)).

4.2.4 Least Trimmed Square Method (lts)

The performance of LMS estimator is not considered good from asymptotic point of view. To remove this drawback, LTS method is used where squared error terms are ordered from smallest to largest. We minimize the sum of first h ordered errors term given as follows

$$\sum_{i=1}^{h}(y_i - \widehat{\beta}_o - \widehat{\beta}_1 x_i)^2,$$

where $h = n/2 + 1$. The statistical properties of LTS are given by Rousseeuw and Leroy (1987).

4.2.5 Least Absolute Deviation (lad)

The method of least absolute deviation was developed by Roger Joseph Boscovich in 1757 with the aim to provide more efficient estimator than usual OLS estimator under the

influence of outliers. In this method, estimates are chosen such that the sum of absolute values of errors is minimized. We minimize the function

$$\min \sum_{i=1}^{n} |\varepsilon_i|.$$

The calculation of LAD estimator is more complex than least square estimators as no exact formula is involved. Considerable amount of outliers in X direction can impact LAD estimators due to the low breaking point of $1/n$ or 0 %.

4.3 Existing Estimators

Ali *et al.* (2021) generalized the traditional ratio type estimator given by Zaman and Bulut (2019) for sensitive setup. They also proposed a new regression type estimator using robust method of estimation, given by

$$z_{r_{\mathrm{i}}} = \bar{z} + b_i(\overline{X} - \bar{x}), \quad \text{for} \quad \text{i = hbm, hmm, hpm, tkm, lms, lts, lad.} \tag{4.3.1}$$

The MSE of the above mentioned estimator is obtained using Taylor series expansion, given by

$$\mathrm{MSE}(z_{r_{\mathrm{i}}}) = \lambda \left[\overline{Z}^2 C_z^2 - 2\beta_i \overline{X}\overline{Z} \rho_{zx} C_x C_z + \beta_i^2 \overline{X}^2 C_x^2 \right], \tag{4.3.2}$$

where $\lambda = \frac{1-f}{n}$ and $f = \frac{n}{N}$. It is proved that the estimator proposed by Ali *et al.* (2021) is more efficient than the generalized ratio type estimator given by Zaman and Bulut (2019). Therefore, any estimator which is more efficient than the estimator given by Ali *et al.* (2021) will automatically perform better than the generalized case estimator. Hence in

this study, we skip the efficiency comparison of proposed estimator with the generalized case and shift our focus to prove that the proposed estimator is better than the estimator given by Ali *et al.* (2021).

4.4 Proposed Class of Estimators

Motivated from Ali *et al.* (2021), we propose a new regression type estimator of the population mean for a sensitive setup. The regression type robust estimator is given by

$$t_{r_i} = k_i\bar{z} + b_i(\overline{X} - \bar{x}), \quad \text{for} \quad \text{i= hbm, hmm, hpm, tkm, s, lms, lts, lad,} \tag{4.4.1}$$

where k_i is appropriate constant to determined and b_i is the sample regression coefficient computed from above mentioned method of estimation discussed in Section 4.2. Using Taylor series expansion, MSE of the suggested estimator is defined as

$$h(\bar{x},\bar{z}) - h(\overline{X},\overline{Z}) = \left[\frac{\partial h(\bar{x},\bar{z})}{\partial \bar{x}}\right]_{\overline{X},\overline{Z}} (\bar{x} - \overline{X}) + \left[\frac{\partial h(\bar{x},\bar{z})}{\partial \bar{z}}\right]_{\overline{X},\overline{Z}} (\bar{z} - \overline{Z}), \tag{4.4.2}$$

where $h(\bar{x},\bar{z}) = t_{r_i}$ and $h(\overline{X},\overline{Z}) = \overline{T}$.

$$h(\bar{x},\bar{z}) - h(\overline{X},\overline{Z}) = \left[\frac{\partial (k_i\bar{z} + b_i(\overline{X} - \bar{x}))}{\partial \bar{x}}\right]_{\overline{X},\overline{Z}} (\bar{x} - \overline{X}) + \left[\frac{\partial (k_i\bar{z} + b_i(\overline{X} - \bar{x}))}{\partial \bar{z}}\right]_{\overline{X},\overline{Z}} (\bar{z} - \overline{Z}). \tag{4.4.3}$$

Now partially differentiating 1^{st} and 2^{nd} terms of right hand side of the Equation (4.4.3) w.r.t $\bar{x}$ and $\bar{z}$ respectively, we get

$$t_{r_i} - \overline{T} = -b_i(\bar{x} - \overline{X}) + k_i(\bar{z} - \overline{Z}). \tag{4.4.4}$$

Now squaring and taking expectation on both the side of the Equation (4.4.4)and then considering $b_i = \beta_i$, where β_i is the population regression coefficient, we get the mean square error of the proposed estimator, given by

$$\begin{aligned}
\text{MSE}(t_{r_\text{i}}) &= E(t_{r_\text{i}} - \overline{T})^2 \\
&= E\left[-b_i(\overline{x} - \overline{X}) + k_i(\overline{z} - \overline{Z})\right]^2 \\
&= E\left[b_i^2(\overline{x} - \overline{X})^2 + k_i^2(\overline{z} - \overline{Z})^2 - 2b_i k_i(\overline{x} - \overline{X})(\overline{z} - \overline{Z})\right] \\
&= \left[\beta_i^2 V(\overline{x}) + k_i^2 V(\overline{z}) - 2\beta_i k_i \text{Cov}(\overline{xz})\right] \\
&= \left(\frac{1-f}{n}\right)\left[\beta_i^2 S_x^2 + k_i^2 S_z^2 - 2\beta_i k_i S_{zx}\right]. \qquad (4.4.5)
\end{aligned}$$

Now, partially differentiating the Equation (4.4.5) with respect to k_i, to obtain the optimum value of k_i that minimizes the MSE of t_{r_i}, we get

$$k_i = \beta_i \frac{S_{zx}}{S_z^2}. \qquad (4.4.6)$$

Substituting the value of k_i in 4.4.5 , the minimum MSE of the proposed estimator is given by

$$\begin{aligned}
\text{MSE}_{\min}(t_{r_\text{i}}) &= \left(\frac{1-f}{n}\right)\left[\beta_i^2 S_x^2 + \beta_i^2 \frac{S_{zx}^2}{S_z^4} S_z^2 - 2\beta_i^2 \frac{S_{zx}^2}{S_z^2}\right] \\
&= \left(\frac{1-f}{n}\right)\beta_i^2\left[S_x^2 - \frac{S_{zx}^2}{S_z^2}\right] \\
&= \left(\frac{1-f}{n}\right)\beta_i^2 S_x^2\left[1 - \frac{S_{zx}^2}{S_z^2 S_x^2}\right] \\
&= \lambda\beta_i^2 \overline{X}^2 C_x^2(1-\rho_{zx}^2), \quad \text{for} \quad \text{i= hbm, hmm, hpm, tkm, s, lms, lts, lad,} \\
& \qquad (4.4.7)
\end{aligned}$$

where C_x is the coefficient of variation and ρ_{zx} is the correlation coefficient between Z and X. Table 4.1 gives the class of estimators which consist of all the above given methods discussed under Section 4.2.

Table 4.1: Family of Proposed Class of Estimators

S. No.	Estimators	MSE
1	$t_{r_{\text{hbm}}} = k\bar{z} + b_{\text{hbm}}(\overline{X} - \bar{x})$	$\lambda \beta_{\text{hbm}}^2 \overline{X}^2 C_x^2 (1 - \rho_{zx}^2)$
2	$t_{r_{\text{hmm}}} = k\bar{z} + b_{\text{hmm}}(\overline{X} - \bar{x})$	$\lambda \beta_{\text{hmm}}^2 \overline{X}^2 C_x^2 (1 - \rho_{zx}^2)$
3	$t_{r_{\text{hpm}}} = k\bar{z} + b_{\text{hpm}}(\overline{X} - \bar{x})$	$\lambda \beta_{\text{hpm}}^2 \overline{X}^2 C_x^2 (1 - \rho_{zx}^2)$
4	$t_{r_{\text{tkm}}} = k\bar{z} + b_{\text{tkm}}(\overline{X} - \bar{x})$	$\lambda \beta_{\text{tkm}}^2 \overline{X}^2 C_x^2 (1 - \rho_{zx}^2)$
5	$t_{r_{\text{s}}} = k\bar{z} + b_{\text{s}}(\overline{X} - \bar{x})$	$\lambda \beta_{\text{s}}^2 \overline{X}^2 C_x^2 (1 - \rho_{zx}^2)$
6	$t_{r_{\text{lms}}} = k\bar{z} + b_{\text{lms}}(\overline{X} - \bar{x})$	$\lambda \beta_{\text{lms}}^2 \overline{X}^2 C_x^2 (1 - \rho_{zx}^2)$
7	$t_{r_{\text{lts}}} = k\bar{z} + b_{\text{lts}}(\overline{X} - \bar{x})$	$\lambda \beta_{\text{lts}}^2 \overline{X}^2 C_x^2 (1 - \rho_{zx}^2)$
8	$t_{r_{\text{lad}}} = k\bar{z} + b_{\text{lad}}(\overline{X} - \bar{x})$	$\lambda \beta_{\text{lad}}^2 \overline{X}^2 C_x^2 (1 - \rho_{zx}^2)$

4.5 Numerical Illustration

In support of theoretical results, we considered two populations with different correlation values. The scrambling variable S is deemed to follow Normal distribution with mean equal to zero and standard deviation equal to 10% of standard deviation of X. The response variable Z is thus reported as $Z = Y + S$. Population 1 is taken from Singh (2003) [pg no-1111], where X denotes the amount of non-real estate farm loans during the year 1977 and Y denotes the amount of real estate farm loans during the year 1977. Population 2 is taken from Singh and Chaudhary (1986) [pg no-177], where X denotes the cultivated area under wheat in a region during the year 1973 and Y denotes the area under wheat in a region during the year 1974. We add the random noise to the existing populations by altering

a certain percentage of the data entries to random values within the data range as it does not modify the value distribution. Assuming the last five observations were incorrectly reported in both the populations. As a result, the addition of outliers contaminates both the real populations with 10% and 12% outliers in the direction of Y. The statistics of the populations under consideration are given in Table 4.2.

Table 4.2: Descriptive Statistics of the Real Populations

Population 1			
$N = 50$	$S_x^2 = 1176526$	$\beta_{\text{hbm}} = 0.4276115$	$\beta_{\text{s}} = 0.3637$
$n = 10$	$S_y^2 = 342021.5$	$\beta_{\text{hmm}} = 0.4140589$	$\beta_{\text{lms}} = 0.3197$
$\overline{X} = 878.1624$	$\rho_{yx} = 0.8038341$	$\beta_{\text{hpm}} = 0.4289556$	$\beta_{\text{lts}} = 0.3521$
$\overline{Y} = 555.4345$	$\beta_{\text{ols}} = 0.434$	$\beta_{\text{tkm}} = 0.4245928$	$\beta_{lad} = 0.4579807$
Population 2			
$N = 34$	$S_x^2 = 22652.05$	$\beta_{\text{hbm}} = 0.9788785$	$\beta_{\text{s}} = 0.948$
$n = 6$	$S_y^2 = 22564.56$	$\beta_{\text{hmm}} = 0.9673564$	$\beta_{\text{lms}} = 1.041$
$\overline{X} = 208.8824$	$\rho_{yx} = 0.9800867$	$\beta_{\text{hpm}} = 0.9813628$	$\beta_{\text{lts}} = 0.9441$
$\overline{Y} = 199.4412$	$\beta_{\text{ols}} = 0.9929$	$\beta_{\text{tkm}} = 0.9567706$	$\beta_{\text{lad}} = 0.9629817$

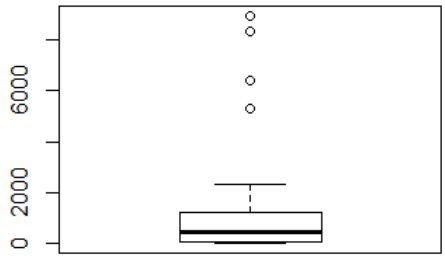

Figure 4.1: Box plot of population 1

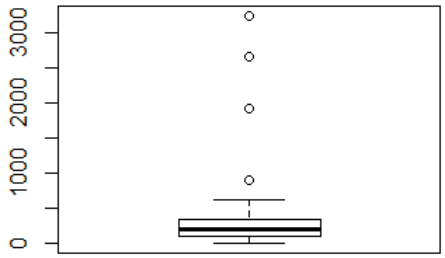

Figure 4.2: Box plot of population 2

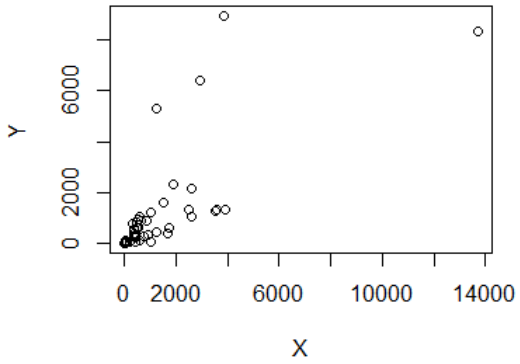

Figure 4.3: Scatter plot of population 1

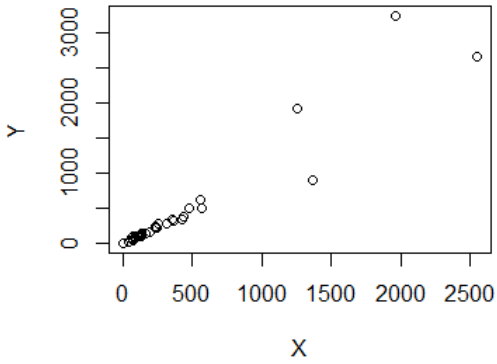

Figure 4.4: Scatter plot of population 2

A box plot is a standardised approach of depicting distribution of the data based on minimum, first quartile, median, third quartile, and maximum observation in the data and can provide information about your outliers and their values. The box plots of the real populations under considerations are given in Figures 4.1 and 4.2 that demonstrate the presence of outliers in the data. The box plots are followed by scatter plots given in Figures 4.3 and 4.4 to show the scatterings between the study variable and auxiliary

variable in the data. Hence, an alternative method of regression is recommended to reduce the effect of outliers. The MSE is calculated using the following Equation (4.5.1)

$$\text{PRE}(z_{r_\text{i}}, t_{r_\text{i}}) = \frac{\text{MSE}(z_{r_\text{i}})}{\text{MSE}(t_{r_\text{i}})} * 100, \quad \text{where} \quad \text{i = hbm, hmm, hpm, tkm, s, lms, lts, lad.} \tag{4.5.1}$$

Table 4.3: MSE and PRE of Estimators for both the real population

	Population 1		Population 2	
Estimators	MSE	PRE	MSE	PRE
z_{r_ls}	152206.5		7713.82	
t_{r_ls}	79815.43	190.70	7126.25	108.24
z_{r_hbm}	154848.1		8475.70	
t_{r_hbm}	59936.46	258.35	5470.65	154.93
z_{r_hmm}	181589.7		8785.17	
t_{r_hmm}	26071.79	696.50	5230.46	167.96
z_{r_hpm}	156913.00		8808.60	
t_{r_hpm}	54231.61	289.34	5213.96	168.94
z_{r_tkm}	190351.7		8777.49	
t_{r_tkm}	20986.99	906.99	5235.91	167.64
z_{r_s}	194271.8		8665.09	
t_{r_s}	19062.61	1019.12	5318.37	162.92
z_{r_lms}	194621.8		9222.80	
t_{r_lms}	18899.58	1029.76	4950.37	186.30
z_{r_lts}	193898.3		158.66	
t_{r_lts}	19238.06	1007.88	5397.53	158.66
z_{r_lad}	156682.5		8688.01	
t_{r_lad}	54783.09	286.00	5301.14	163.88

Table 4.4: MSE and PRE of Estimators for both the real population

	Population 1		**Population 2**	
Estimators	**MSE**	**PRE**	**MSE**	**PRE**
$z_{r_{\text{hbm}}}$	10637.36		150.5338	
$t_{r_{\text{hbm}}}$	6408.435	165.9899	143.1868	105.131
$z_{r_{\text{hmm}}}$	10676.85		150.9738	
$t_{r_{\text{hmm}}}$	6008.658	177.6911	139.8358	107.965
$z_{r_{\text{hpm}}}$	10635.32		150.5471	
$t_{r_{\text{hpm}}}$	6448.785	164.9198	143.9145	104.6087
$z_{r_{\text{tkm}}}$	10643.16		2415.272	
$t_{r_{\text{tkm}}}$	6318.274	168.4504	136.7921	111.1948
$z_{r_{\text{s}}}$	11070.78		153.5714	
$t_{r_{\text{s}}}$	4743.652	233.3809	134.2957	114.3531
$z_{r_{\text{lms}}}$	11910.2		162.385	
$t_{r_{\text{lms}}}$	**3582.106**	**332.4915**	161.9373	100.2764
$z_{r_{\text{lts}}}$	11297.76		154.3767	
$t_{r_{\text{lts}}}$	4344.955	260.0201	**133.193**	**115.9045**
$z_{r_{\text{lad}}}$	10674.41		151.3571	
$t_{r_{\text{lad}}}$	7351.019	145.2099	138.5739	109.2248

Table 4.4 contain MSE and PRE of the proposed class of estimators as compared to Ali *et al.* (2021) generalized estimators. Results show that the proposed class of estimators have less MSE for all the robust cases. For Population 1, t_{lms} attains the least value of MSE among all the stated robust cases. For Population 2, t_{lts} has the minimum value of MSE amongst all proposed robust estimators.

Table 4.4 contain MSE and PRE of the proposed class of estimators as compared to Ali *et al.* (2021) generalized estimators. Results show that the proposed class of estimators have less MSE for all the robust cases. For both the real populations, t_{lms} attains the least value of MSE among all the stated robust cases.

4.6 Simulation Study

A simulation model is set up to inspect the dynamics of robust methods of estimation more precisely. Two artificial bivariate populations are considered with different mean and covariance matrix for the comparison of the mean square errors of the proposed estimator with the competing estimators. Population 1 is an Artificial Population (AP1) taken from a bivariate normal distribution with pre define mean vector and variance-covariance matrix given by

$$\mu = [3,3] \qquad \text{and} \quad \Sigma = \begin{bmatrix} 12 & 3 \\ 3 & 6 \end{bmatrix}, \quad \text{respectively.}$$

Population 2 is Artificial Population (AP2) generated by considering the real parameters of the Population 1 used in numerical example of Section 5. The mean and covariance matrix of AP2 are given by

$$\mu = [555.4345, 878.1624] \quad \text{and} \quad \Sigma = \begin{bmatrix} 11765263 & 509910.5 \\ 509910.5 & 342021.5 \end{bmatrix}, \quad \text{respectively.}$$

We introduce noise to the Y of the artificial population to make it suitable for evaluating the performance of proposed estimator in the presence of 2% and 1% outliers in Artificial Population 1 and 2 respectively. Figures 4.5 and 4.6 demonstrate the box plots of artificial data derived from the bivariate normal populations which are also followed by a scatter plots of both the populations given in Figures 4.7 and 4.8. It is evident from the diagrams that data contain outliers and hence an alternative method is required to gain precision using robust techniques.

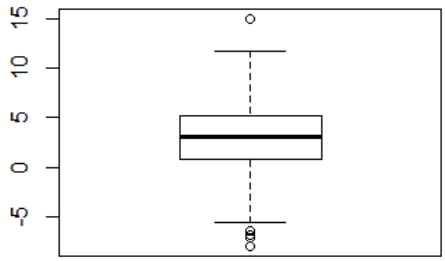

Figure 4.5: Box plot of AP1

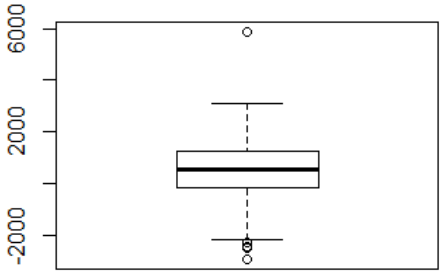

Figure 4.6: Box plot of AP2

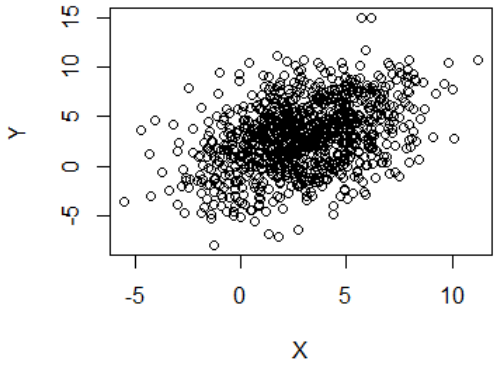

Figure 4.7: Scatter plot of AP1

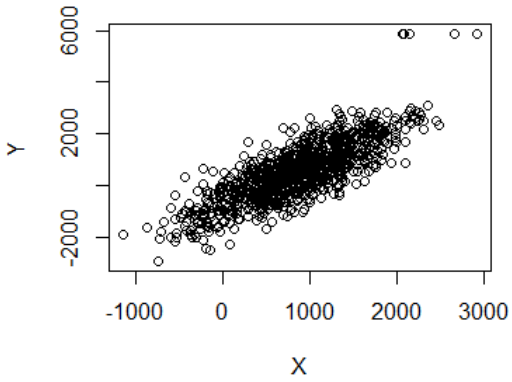

Figure 4.8: Scatter plot of AP2

Table 4.6 contains the MSEs and PREs for AP1 and AP2 respectively. The Proposed class of estimator has lesser MSE than the generalized estimator stated by Ali *et al.* (2021). For both the Artificial Populations, t_{lts} attains minimum MSE. Thus, provides greatest efficiency among the entire class of robust estimators.

Table 4.5: MSE and PRE of Estimators for Artificial Populations

	Population 1		Population 2	
Estimators	**MSE**	**PRE**	**MSE**	**PRE**
$z_{r_{\text{hbm}}}$	0.1828978		7417.141	
$t_{r_{\text{hbm}}}$	0.02921197	626.1058	4912.94	150.9715
$z_{r_{\text{hmm}}}$	0.1828948		7417.071	
$t_{r_{\text{hmm}}}$	0.02908553	628.8172	4908.575	151.1044
$z_{r_{\text{hpm}}}$	0.1828833		7417.006	
$t_{r_{\text{hpm}}}$	0.02831145	645.9693	4904.136	151.2398
$z_{r_{\text{tkm}}}$	0.1828938		7417.077	
$t_{r_{\text{tkm}}}$	0.04077999	629.8	4908.958	151.0927
$z_{r_{\text{s}}}$	0.1843017		7419.404	
$t_{r_{\text{s}}}$	0.04077999	451.9415	4992.912	148.5987
$z_{r_{\text{lms}}}$	0.18687		7439.285	
$t_{r_{\text{lms}}}$	0.05079775	367.8705	7439.285	166.242
$z_{r_{\text{lts}}}$	0.1844158		7469.268	
$t_{r_{\text{lts}}}$	0.01721933	**1070.981**	4280.854	**174.4808**
$z_{r_{\text{lad}}}$	0.1829533		7417.252	
$t_{r_{\text{lad}}}$	0.03063507	597.2021	4794.333	154.7087

4.7 Results and Discussion

Table 4.1 demonstrates the family of a proposed class of estimators with their respective MSE, discussed under Section 4. MSE of different estimators are obtained using the Taylor series expansion. The theoretical results are verified with the help of numerical study using two real and artificial populations. Table 4.2 gives the parametric information of the real data that are used in the numerical study. Tables 4.4 and 4.6 showcase numerical estimates of MSE and PRE of the proposed and competing class of estimators for both the populations under real and artificial datasets respectively. From the results of the Table 4.4 we can infer that the proposed estimator under the setting of least median square has

Table 4.6: MSE and PRE of Estimators for Artificial Populations

	Population 1		Population 2	
Estimators	MSE	PRE	MSE	PRE
$z_{r_{\text{hbm}}}$	0.1847		8331.92	
$t_{r_{\text{hbm}}}$	0.0292	632.53	5393.58	154.33
$z_{r_{\text{hmm}}}$	0.1847		8343.48	
$t_{r_{\text{hmm}}}$	0.0288	639.49	5000.91	166.83
$z_{r_{\text{hpm}}}$	0.1847		8333.96	
$t_{r_{\text{hpm}}}$	0.0283	652.62	5108.01	163.15
$z_{r_{\text{tkm}}}$	0.1847		8343.47	
$t_{r_{\text{tkm}}}$	0.0288	640.26	5000.98	166.83
$z_{r_{\text{s}}}$	0.1861		8324.20	
$t_{r_{\text{s}}}$	0.0407	456.40	5386.31	154.54
$z_{r_{\text{lms}}}$	0.1886		8357.40	
$t_{r_{\text{lms}}}$	0.0508	371.360	5891.94	141.84
$z_{r_{\text{lts}}}$	0.1861		8451.87	
$t_{r_{\text{lts}}}$	0.0180	1030.82	4436.49	190.50
$z_{r_{\text{lad}}}$	0.1848		8331.25	
$t_{r_{\text{lad}}}$	0.0306	603.31	5147.87	161.83

observed maximal value of PRE as 1029.76 and 186.30 for the real populations 1 and 2 respectively. Equivalently, from the results provided in Table 4.6, we observed that the maximum gain in efficiency is achieved by the proposed estimator under the setting of least trimmed square with maximum PRE value of 1030.82 and 190.50 for AP1 and AP2 respectively.

4.8 Conclusion

When dataset is contaminated by one or a few outliers, finding such observations becomes a severe problem. We observe that most data sets have more outliers or a cluster of in-

fluential occurrences. The robust estimating approach is a type of regression analysis that is intended to overcome some of the limitations of classic parametric and non-parametric methods. Robust regression is less susceptible to the effects of outliers and can produce more accurate results in comparison to the conventional OLS.

From the above study, it is evident that the proposed estimator in which different alternative methods of regression like LTS, LMS, LAD, Tukey M, Hampel M, Huber MM and S method of estimation are involved is more efficient than estimator given by Ali *et al.* (2021). Since Ali *et al.* (2021) also generalized Zaman and Bulut (2019) estimators for sensitive setup and proved that their proposed estimator are more efficient than the generalized estimator given by Zaman and Bulut (2019). Hence our proposed estimator is also better than the generalized estimator given by Zaman and Bulut (2019). Therefore, it is recommended to use the proposed family of estimator in a sensitive setup under the alternative method of regression when data contain outliers.

Chapter 5

Generalized Double Sampling Family of Estimators for Population mean of Sensitive Variable Harnessing Non-sensitive Auxiliary Parameter

5.1 Introduction

The Randomized Response Technique (RRT) addresses the need that frequently emerges in areas of socio-economics, cognitive, and health research for elements of exceedingly sensitive subjects. In most surveys, the interviewee is hesitant to discuss their past experiences with drugs, medical illnesses, sexual conduct, abortions, and so on. In the following chapter, we introduce a generalized double sampling family of estimator to estimate the population mean of the sensitive variable using a non sensitive supplementary variable. The remainder of the chapter is divided into the following sections: In Section 2, we inspect several estimators of the finite population mean that are accessible in litera-

ture. Section 3 provides the bias and Mean Squared Error (MSE) of the suggested family of estimators derived using the Taylor series approximation, combined with the optimal requirement and minimal MSE. The theoretical efficiency comparisons of the suggested estimators with competing estimators are presented in Section 4, and the efficiency criteria over competing estimators is derived. To assess the performances of the members of the mentioned classes of estimators, an observational research is conducted in Section 5. A simulation analysis is conducted in Section 6 to demonstrate the effectiveness of various members of the suggested and competing classes of estimators. Section 7 presents the results of the investigation as well as a discussion. Finally, the conclusion derived from the results are offered in Section 8.

5.2 Some Existing Estimators

Let Y be the research variable, which comprises delicate features that may not be accessed exactly as a result of the respondent's response. Let X be a non-sensitive secondary variate that is correlated positively to Y. Let S be a scrambled variate with a predefined distribution which is uncorrelated with Y and X. The sensitive variable Y is scrambled using the variable S. To assess Y, each respondent is instructed to choose a random number from the S distribution, say s, and add it to the actual value of Y. Let Z denote the reported scrambled response to Y that was initially suggested by Warner 1965 and elaborated by Pollock and Bek 1976, given by

$$Z = Y + S.$$

In the double sampling or two-phase sampling scheme, first a preliminary large sample of size n' (referred to as the first phase sample) is taken from a population of size N, and then a sub sample of size n (referred to as the second phase sample) is drawn from the first phase sample using a simple random sampling without replacement scheme at both places.

Let y_i and x_i, $(i = 1,2,\ldots,n)$ represent the sample values of the research and non-sensitive supplementary variables on the i^{th} units, respectively. When the characteristic provided on secondary variable, X is not taken into account, the typical RRT sample mean is considered to be an unbiased estimator of the population mean of a sensitive variable, stated as

$$\mu_0 = \bar{z}. \tag{5.2.1}$$

The variance of μ_0 is given by

$$\text{MSE}(\mu_0) = \lambda_n(S_y^2 + S_s^2),$$

where

$$\lambda_n = \left(\frac{1}{n} - \frac{1}{N}\right), \quad S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2 \quad \text{and} \quad S_s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(s_i - \overline{S})^2.$$

Using the available non-sensitive supplementary data, Sousa *et al.* 2010 improved the RRT estimator of the population mean of the sensitive variable by introducing ratio estimator under simple random sampling scheme. Under double sampling, Saleem *et al.* 2019 proposed a ratio estimator of a population mean of a susceptible research variate,

given by

$$\mu_R = \overline{z}\left(\frac{\overline{x}'}{\overline{x}}\right) \tag{5.2.2}$$

This estimator is biased for the population mean of the susceptible research variate. The bias and MSE of the aforementioned estimator, accurate up to the first degree of approximation is as follows

$$\text{MSE}(\mu_R) = \lambda_n \overline{Z}^2 \left(C_z^2 + C_x^2 - 2\rho_{zx} C_z C_x\right) - \lambda_n' \overline{Z}^2 \left(C_x^2 - 2\rho_{zx} C_z C_x\right) \tag{5.2.3}$$

where

$$\lambda_n' = \left(\frac{1}{n'} - \frac{1}{N}\right), \quad S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2, \quad C_x^2 = \frac{S_x^2}{\overline{X}^2}, \quad C_y^2 = \frac{S_y^2}{\overline{Y}^2}, \quad \text{and} \quad C_z^2 = C_y^2 + C_s^2,$$

and ρ_{zx} is the correlation coefficient between the variable Z and X. Gupta *et al.* 2017 further modified the ratio estimator and introduced ratio exponential estimator under simple random sampling scheme. The modified estimator under double sampling scheme is given by

$$\mu_{ER} = \overline{z} \exp\left(\frac{\overline{x}' - \overline{x}}{\overline{x}' + \overline{x}}\right). \tag{5.2.4}$$

The MSE of the modified ratio estimator μ_{ER} is given by

$$\text{MSE}(\mu_{ER}) = \lambda_n \overline{Z}^2 \left(C_z^2 + \frac{1}{4}C_x^2 - 2\rho_{zx} C_z C_x\right) - \lambda_n' \overline{Z}^2 \left(\frac{1}{4}C_x^2 - 2\rho_{zx} C_z C_x\right). \tag{5.2.5}$$

5.3 Proposed Generalized Class of Estimators

A new family of generalized estimators for the finite population mean of the sensitive research variate employing a single non-sensitive supplementary variate is introduced in this section. Under first order of approximation, expressions for the bias and mean squared error of the suggested family of estimators are produced. Motivated from Misra and Singh (2016), a generalized two-phase sampling estimator $\bar{z}_d$, expressed as the bounded function of $(\bar{z},\bar{x},\bar{x}')$ is given as

$$\bar{z}_d = g(\bar{z},\bar{x},\bar{x}'), \tag{5.3.1}$$

meeting the required condition given by Taylors series expansion, such that

1. $g(\overline{Z},\overline{X},\overline{X}) = \overline{Z}$,

2. first order partial differential coefficient of $g(\bar{z},\bar{x},\bar{x}')$ with respect to $\bar{z}$ at $T = (\overline{Z},\overline{X},\overline{X})$ is unity, i.e. $g_0 = \left[\frac{\partial}{\partial \bar{z}} g(\bar{z},\bar{x},\bar{x}')\right]_T = 1$,

3. $g_{00} = \left[\frac{\partial^2}{\partial \bar{z}^2} g(\bar{z},\bar{x},\bar{x}')\right]_T = 0$,

4. $g_1 = -g_2$,

   where g_1 and g_2 being the first order partial derivative of $g(\bar{z},\bar{x},\bar{x}')$ with respect to $\bar{x}$ and $\bar{x}'$ respectively at the point $T = (\overline{Z},\overline{X},\overline{X})$, i.e.

$$g_1 = \left[\frac{\partial}{\partial \bar{x}} g(\bar{z},\bar{x},\bar{x}')\right]_T \quad \text{and} \quad g_2 = \left[\frac{\partial}{\partial \bar{x}'} g(\bar{z},\bar{x},\bar{x}')\right]_T,$$

5. $g_{01} = -g_{02}$, where

$$g_{01} = \left[\frac{\partial^2}{\partial\bar{z}\partial\bar{x}} g(\bar{z},\bar{x},\bar{x}')\right]_T \quad \text{and} \quad g_{02} = \left[\frac{\partial^2}{\partial\bar{z}\partial\bar{x}'} g(\bar{z},\bar{x},\bar{x}')\right]_T.$$

To estimate the bias and mean squared error of $\bar{z}_d$, we consider the following assumption

$$\bar{z} = \overline{Z}(1+e_0),\ \bar{x} = \overline{X}(1+e_1),\ \bar{x}' = \overline{X}(1+e_1'),$$

such that

$$E(e_0) = E(e_1) = E(e_1') = 0,$$

$$E(e_0^2) = \lambda C_z^2,\ E(e_1^2) = \lambda C_x^2,\ E(e_1'^2) = \lambda' C_x^2,$$

$$E(e_0e_1) = \lambda\rho_{zx}C_zC_x,\ E(e_0e_1') = \lambda'\rho_{zx}C_zC_x,\ E(e_1e_1') = \lambda' C_x^2.$$

Expressing $g(\bar{z},\bar{x},\bar{x}')$ in the third order Taylors series expansion about the point $T = (\overline{Z},\overline{X},\overline{X})$, we have

$$\begin{aligned}\bar{z}_d = {} & g(\overline{Z},\overline{X},\overline{X}) + (\bar{z}-\overline{Z})g_0 + (\bar{x}-\overline{X})g_1 + (\bar{x}'-\overline{X})g_2 + \frac{1}{2!}\left\{(\bar{z}-\overline{Z})^2 g_{00} + (\bar{x}-\overline{X})^2 g_{11}\right. \\ & \left. + (\bar{x}'-\overline{X})^2 g_{22} + 2(\bar{z}-\overline{Z})(\bar{x}-\overline{X})g_{01} + 2(\bar{z}-\overline{Z})(\bar{x}'-\overline{X})g_{02} + 2(\bar{x}-\overline{X})(\bar{x}'-\overline{X})g_{12}\right\} \\ & + \frac{1}{3!}\left\{\frac{\partial}{\partial\bar{z}} + \frac{\partial}{\partial\bar{x}} + \frac{\partial}{\partial\bar{x}'}\right\}^3 g(\bar{z}^*,\bar{x}^*,\bar{x}'^*), \end{aligned} \tag{5.3.2}$$

where

$$\bar{z}^* = \overline{Z} + h(\bar{z}-\overline{Z}), \quad \bar{x}^* = \overline{X} + h(\bar{x}-\overline{X}) \quad \text{and} \quad \bar{x}'^* = \overline{X} + h(\bar{x}'-\overline{X}), \quad 0 < h < 1,$$

and $g_0, g_1, g_2, g_{00}, g_{01}, g_{02}$ are defined earlier, also

$$g_{11} = \left[\frac{\partial^2}{\partial \overline{x}^2} g(\overline{z}, \overline{x}, \overline{x}')\right]_T, \quad g_{22} = \left[\frac{\partial^2}{\partial \overline{x}'^2} g(\overline{z}, \overline{x}, \overline{x}')\right]_T \text{ and } \quad g_{12} = \left[\frac{\partial}{\partial \overline{x} \partial \overline{x}'} g(\overline{z}, \overline{x}, \overline{x}')\right]_T.$$

Now using the conditions discussed on page no.87 earlier, we have

$$\begin{aligned}\overline{z}_d - \overline{Z} =& \overline{Z}e_0 + \overline{X}e_1 g_1 - \overline{X}e_1' g_1 + \frac{1}{2!}\left\{\overline{X}^2 e_1^2 g_{11} + \overline{X}^2 e_1'^2 g_{22} + 2\overline{ZX} e_0 e_1 g_{01} - 2\overline{ZX} e_0 e_1' g_{01}\right. \\ &\left. + 2\overline{X}^2 e_1 e_1' g_{12}\right\} + \frac{1}{3!}\left\{\frac{\partial}{\partial \overline{z}} + \frac{\partial}{\partial \overline{x}} + \frac{\partial}{\partial \overline{x}'}\right\}^3 g(\overline{z}^*, \overline{x}^*, \overline{x}'^*). \end{aligned} \tag{5.3.3}$$

Taking expectation on both the sides of the above equation, we get the bias of the proposed family of estimators up to the first order of approximation, given by

$$\text{Bias}(\overline{z}_d) = \frac{1}{2!}\left\{\overline{X}^2\left(\lambda_n g_{11} + \lambda_n' g_{22} + 2\lambda_n' g_{12}\right) C_x^2 + 2\overline{ZX} C_z C_x \rho_{zx}(\lambda_n - \lambda_n') g_{01}\right\}. \tag{5.3.4}$$

Squaring both the side of the Equation (5.3.3) and then taking expectations, we get MSE of the suggested family of estimators up to the first order of approximation, given by

$$\begin{aligned}\text{MSE}(\overline{z}_d) &= E(\overline{z}_d - \overline{Z})^2 \\ &= E\left\{\overline{Z}e_0 + \overline{X}e_1 g_1 - \overline{X}e_1' g_1\right\}^2 \\ &= \overline{Z}^2 E(e_0^2) + \overline{X}^2 g_1^2 E(e_1^2) + \overline{X}^2 g_1^2 E(e_1'^2) - 2\overline{X}^2 g_1^2 E(e_1 e_1') + 2\overline{ZX} g_1 E(e_0 e_1) - 2\overline{ZX} g_1 E(e_0 e_1'). \end{aligned} \tag{5.3.5}$$

Using the value of the different expectations in Equation (5.3.5), we get

$$\begin{aligned}\text{MSE}(\overline{z}_d) &= \overline{Z}^2 \lambda_n C_z^2 + \overline{X}^2 \lambda_n C_x^2 g_1^2 + \overline{X}^2 \lambda_n' C_x^2 g_1^2 - 2\overline{X}^2 \lambda_n' C_x^2 g_1^2 + 2\overline{ZX} \lambda_n \rho_{zx} C_z C_x g_1 - 2\overline{ZX} \lambda_n' \rho_{zx} C_z C_x g_1 \\ &= \lambda_n \overline{Z}^2 C_z^2 + \overline{X}^2 C_x^2 (\lambda_n - \lambda_n') g_1^2 + 2\overline{ZX}(\lambda_n - \lambda_n') \rho_{zx} C_z C_x g_1 \\ &= \lambda_n \overline{Z}^2 C_z^2 + (\lambda_n - \lambda_n')\left[\overline{X}^2 C_x^2 g_1^2 + 2\overline{ZX} \rho_{zx} C_z C_x g_1\right]. \end{aligned} \tag{5.3.6}$$

Minimizing Equation (5.3.6) to get the optimal value of unknown constant g_1, we get

$$g_1\overline{X}^2C_x^2+\overline{ZX}\rho_{zx}C_zC_x=0,$$

$$g_1(\text{optimum})=\frac{-\overline{Z}\rho_{zx}C_z}{\overline{X}C_x}.$$

Putting the value of g_1 in Equation (5.3.6), we get the minimum value of MSE as

$$\text{MSE}_{\min}(\bar{z}_d)=\lambda_n\overline{Z}^2C_z^2+(\lambda_n-\lambda_n')\left[\overline{X}^2C_x^2\left(\frac{-\overline{Z}\rho_{zx}C_z}{\overline{X}C_x}\right)^2-2\overline{ZX}\rho_{zx}C_zC_x\left(\frac{-\overline{Z}\rho_{zx}C_z}{\overline{X}C_x}\right)\right],$$

$$\text{MSE}_{\min}(\bar{z}_d)=\lambda_n\overline{Z}^2C_z^2-\left(\frac{1}{n}-\frac{1}{n'}\right)\left[\overline{Z}^2\rho_{zx}^2C_z^2\right]. \tag{5.3.7}$$

Some members of $(\bar{z}_d)$ family of estimators are given by

1. $\bar{z}_{d(1)}=\bar{z}+k_1(\bar{x}-\bar{x}')$,
2. $\bar{z}_{d(2)}=\bar{z}\left(\frac{\bar{x}}{\bar{x}'}\right)\left[1+k_1\left(\frac{\bar{x}}{\bar{x}'}-1\right)\right]$,
3. $\bar{z}_{d(3)}=k_1\bar{z}\left(\frac{\bar{x}}{\bar{x}'}\right)$,
4. $\bar{z}_{d(3)}=\bar{z}\left(\frac{\bar{x}}{\bar{x}'}\right)^{k_1}$,
5. $\bar{z}_{d(4)}=k_1\bar{z}\exp\left(\frac{\bar{x}-\bar{x}'}{\bar{x}+\bar{x}'}\right)$,
6. $\bar{z}_{d(4)}=[\bar{z}+k_1(\bar{x}-\bar{x}')]\exp\left(\frac{\bar{x}-\bar{x}'}{\bar{x}+\bar{x}'}\right)$, [Saleem *et al.* 2019]
7. $\bar{z}_{d(4)}=k_1\bar{z}\left(\frac{\bar{x}}{\bar{x}'}\right)\exp\left(\frac{\bar{x}-\bar{x}'}{\bar{x}+\bar{x}'}\right)$.

These members can easily be demonstrated to satisfy the requirements of $\bar{z}_d$ and thus corresponds to the $\bar{z}_d$ class of estimators. Let us consider the estimator $\bar{z}_{d(1)}$, given by

$$\bar{z}_{d(1)}=\bar{z}+k_1(\bar{x}-\bar{x}'),$$

To obtain the corresponding bias and MSE of $\bar{z}_{d(1)}$, we have

$$\bar{z}_{d(1)} = \overline{Z}(1+e_0) + k_1\left[\overline{X}(1+e_1) - \overline{X}(1+e_1')\right],$$

$$\bar{z}_{d(1)} - \overline{Z} = \overline{Z}e_0 + k_1\overline{X}(e_1 - e_1'). \tag{5.3.8}$$

Taking expectation on both the sides of the above equation, the bias of $\bar{z}_{d(1)}$ is given by

$$\begin{aligned}\text{Bias}(\bar{z}_{d(1)}) &= E(\bar{z}_{d(1)}) - \overline{Z} \\ &= \overline{Z}E(e_0) + k_1\overline{X}(E(e_1) - E(e_1')).\end{aligned}$$

Using the values of expectations, we have $\text{Bias}(\bar{z}_{d(1)}) = 0$. Now squaring Eq. (5.3.8) on both the sides and then taking expectations, the MSE of $\bar{z}_{d(1)}$ is given by

$$\begin{aligned}\text{MSE}(\bar{z}_{d(1)}) &= E(\bar{z}_{d(1)} - \overline{Z})^2 \\ &= E[\overline{Z}e_0 + k_1\overline{X}(e_1 - e_1')]^2 \\ &= E[\overline{Z}e_0 + \overline{X}e_1k_1 - \overline{X}e_1'k_1]^2 \\ &= \overline{Z}^2E(e_0^2) + \overline{X}^2k_1^2E(e_1^2) + \overline{X}^2k_1^2E(e_1') - 2\overline{X}^2k_1^2E(e_1e_1') + 2\overline{ZX}E(e_0e_1) - 2\overline{ZX}k_1E(e_0e_1') \\ &= \overline{Z}^2\lambda_nC_z^2 + \overline{X}^2\lambda_nk_1^2 + \overline{X}^2\lambda_n'C_x^2k_1^2 - 2\overline{X}^2\lambda_n'C_x^2k_1^2 + 2\overline{ZX}\lambda_n\rho_{zx}C_zC_xk_1 - 2\overline{ZX}\lambda_n'\rho_{zx}C_zC_xk_1 \\ &= \overline{Z}^2\lambda_nC_z^2 + \overline{X}^2C_x^2(\lambda_n - \lambda_n')k_1^2 + 2\overline{ZX}(\lambda_n - \lambda_n')\rho_{zx}C_zC_xk_1 \\ &= \overline{Z}^2\lambda_nC_z^2 + (\lambda_n - \lambda_n')\left[\overline{X}^2C_x^2k_1^2 + 2\overline{ZX}\rho_{zx}C_zC_xk_1\right].\end{aligned}$$

The minimum value of MSE $\bar{z}_{d(1)}$ is obtained for the optimum value of k_1 is given by

$$k_1(\text{optimum}) = \frac{-\overline{Z}\rho_{zx}C_z}{\overline{X}C_x},$$

and the MSE of $\bar{z}_{d(1)}$ under the optimum value of the defining scalar is given by

$$\text{MSE}(\bar{z}_{d(1)}) = \lambda_n \overline{Z}^2 C_z^2 - \left(\frac{1}{n} - \frac{1}{n'}\right)\left[\overline{Z}^2 \rho_{zx}^2 C_z^2\right], \tag{5.3.9}$$

which is similar to the least MSE of suggested generalized class of estimator $\bar{z}_d$. Similar verification can also be done for other estimators.

5.4 Efficiency Comparison

In the current section, we theoretically compare the suggested class of estimators against competing estimators of population mean of the sensitive research variate utilizing the pre-define supplementary variable.

1. Double Sampling Ratio Estimator: Using Eq. (5.2.3) and (5.3.7), we have

$$\begin{aligned}
&\text{MSE}(\mu_R) - \text{MSE}(\bar{z}_d) \\
&= \lambda_n \overline{Z}^2 (C_z^2 + C_x^2 - 2\rho_{zx} C_z C_x) - \lambda_n' \overline{Z}^2 (C_x^2 - 2\rho_{zx} C_z C_x) - \lambda_n \overline{Z}^2 C_z^2 + \left(\lambda_n - \lambda_n'\right)\left(\overline{Z}^2 \rho_{zx}^2 C_z^2\right) \\
&= \lambda_n \overline{Z}^2 \left(C_x^2 - 2\rho_{zx} C_z C_x + \rho_{zx}^2 C_z^2\right) - \lambda_n' \overline{Z}^2 \left(C_x^2 - 2\rho_{zx} C_z C_x + \rho_{zx}^2 C_z^2\right) \\
&= \lambda_n \overline{Z}^2 \left(C_x - \rho_{zx} C_z\right)^2 - \lambda_n' \overline{Z}^2 \left(C_x - \rho_{zx} C_z\right)^2 \\
&= \overline{Z}^2 \left(\lambda_n - \lambda_n'\right) \left(C_x - \rho_{zx} C_z\right)^2 \\
&\geq 0,
\end{aligned}$$

which indicates that the suggested generalised double sampling estimator surpasses the RRT ratio estimator.

2. Double Sampling Exponential Ratio Estimator: Using Eq. (5.2.5) and (5.3.7), we have

$$
\begin{aligned}
&\mathrm{MSE}(\mu_{TR}) - \mathrm{MSE}(\bar{z}_d) \\
&= \lambda_n \overline{Z}^2 \left(C_z^2 + \frac{1}{4} C_x^2 - \rho_{zx} C_z C_x \right) - \lambda_n' \overline{Z}^2 \left(\frac{1}{4} C_x^2 - \rho_{zx} C_z C_x \right) - \lambda_n \overline{Z}^2 C_z^2 + \left(\lambda_n - \lambda_n' \right) \left(\overline{Z}^2 \rho_{zx}^2 C_z^2 \right) \\
&= \lambda_n \overline{Z}^2 \left(\frac{1}{4} C_x^2 - \rho_{zx} C_z C_x + \rho_{zx}^2 C_z^2 \right) - \lambda_n' \overline{Z}^2 \left(\frac{1}{4} C_x^2 - \rho_{zx} C_z C_x + \rho_{zx}^2 C_z^2 \right) \\
&= \lambda_n \overline{Z}^2 \left(\frac{1}{2} C_x - \rho_{zx} C_z \right)^2 - \lambda_n' \overline{Z}^2 \left(\frac{1}{2} C_x - \rho_{zx} C_z \right)^2 \\
&= \overline{Z}^2 \left(\lambda_n - \lambda_n' \right) \left(\frac{1}{2} C_x - \rho_{zx} C_z \right)^2 \\
&\geq 0,
\end{aligned}
$$

which indicates that the suggested generalised double sampling estimator surpasses the RRT modified ratio exponential estimator.

5.5 Numerical Illustration

Two real populations are considered to verify the results provided in Section 3. Population 1 is taken from Sarndal and Wretman (1986), where X denotes the population during 1985 and Y denotes the revenue from the 1985 municipal taxation. Population 2 is taken from Murthy (1967), where X denotes the number of workers in the factory and Y denotes the output for 80 factories in a region. The study variable is scrambled using the additive model where the scrambled variable is assumed to follow normal distribution with predefined parameters. The parametric specifications of both the populations under consideration are provided in Table 5.1 given below:

Table 5.1: Descriptive Statistics of the Real Populations

S. No.	Information	Population 1	Population 2
1	N	250	50
2	n'	50	20
3	n	20	5
4	$\overline{Y}$	30.944	555.4345
5	$\overline{X}$	259.276	878.1624
6	S_x	630.8737	1084.678
7	S_y	54.28762	584.826
8	S_z	630.8737	587.1704
9	ρ_{zx}	0.6476499	0.7922381

With a view to analyze the competency of the proposed families of estimators, we have assessed the Percent Relative Efficiencies (PREs) of all the competing estimators under consideration for both the populations. PREs for two proposed families of estimators in comparison to RRT usual mean estimator is determined using the following equation:

$$\text{PRE}(\cdot, \mu_0) = \frac{MSE(\mu_0)}{MSE(\cdot)} \times 100 \tag{5.5.1}$$

Table 5.2 demonstrates the effectiveness of the proposed generalized class of esti-

Table 5.2: MSE and PRE of proposed and competing estimators

Estimators	MSE	PRE	MSE	PRE
μ_0	135.7930	100.0000	61551.84	100.0000
μ_R	70.1455	193.5876	35410.41	173.8242
μ_{ER}	60.33562	225.0627	30840.54	199.5810
$\bar{z}_d$	54.04324	251.2673	28418.01	216.5945

mators over the generalized ratio estimator for both the real populations and infer that the proposed class of estimators is the most efficient among all the estimators for both the population 1 and 2 respectively.

5.6 Simulation Analysis

Under this section, we analyze the effectiveness of the suggested generalized family of estimators with the traditional unbiased RRT estimator, generalized ratio and exponential estimators with a prototypical simulation study using the RRT model proposed by Pollock and Bek (1976). We generated six populations namely Normal, Log normal, Beta, Gamma, Poisson and Uniform, following Singh *et al.* (1998). The study and supplementary variates are generated using the model $Y = 5 + \sqrt{(1-\rho_{xy}^2)}Y^* + \rho_{xy}\left(\frac{S_y}{S_x}\right)X^*$ and $X = 10 + X^*$, where X^* and Y^* are independent variates of corresponding parent distributions. The suggested transformation employs the known correlation coefficient between the research variable Y and the auxiliary character X. For each bivariate population of size $N = 1000$, we consider the sample at first-phase and second-phase of sizes $(50,20),(100,40),(200,80)$ and $(300,120)$ respectively, for different values of correlation coefficient $\rho_{xy} = 0.6, 0.7, 0.8$ and 0.9 respectively. The Percent Relative Efficiency of the suggested estimator in comparison to the randomized response mean estimator for the six simulated populations are given in Tables 5.3-5.8.

Table 5.3: PRE of z_d family of estimators for Normal

$N = 1000$ $X^* \sim N(50,4)$		$Y^* \sim N(40,3)$		$S^* \sim N(0,1)$	
	Estimators	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 50$	μ_R	85.2575	96.65884	107.3416	142.9953
$n = 20$	μ_{ER}	128.7815	141.8115	139.8019	176.3405
	$\bar{z}_d$	129.8607	141.8683	158.5283	182.8886
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 100$	μ_R	84.9889	96.586	114.1999	144.3004
$n = 40$	μ_{ER}	129.538	143.0425	159.6888	179.2076
	$\bar{z}_d$	130.653	143.1016	160.4806	186.1209
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 200$	μ_R	84.42225	96.4312	114.9072	147.1617
$n = 80$	μ_{ER}	131.1798	145.7379	163.9333	185.6379
	$\bar{z}_d$	132.3734	145.8019	164.8044	193.403
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 300$	μ_R	83.81265	96.26288	115.689	150.4155
$n = 120$	μ_{ER}	133.0189	148.7965	168.8288	193.2006
	$\bar{z}_d$	134.3027	148.8663	169.795	202.026

Table 5.4: PRE of z_d family of estimators for lognormal

$X^* \sim LN(6,4)$		$Y^* \sim N(10,3)$		$S^* \sim N(5,3)$	
	Estimators	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 50$	μ_R	63.33768	74.76202	94.42444	134.4934
$n = 20$	μ_{ER}	120.1368	138.454	162.8191	192.0772
	$\bar{z}_d$	127.1376	141.5904	162.9589	196.5392
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 100$	μ_R	62.85155	74.36603	94.31352	135.4813
$n = 40$	μ_{ER}	120.6666	139.6042	165.0248	195.9053
	$\bar{z}_d$	127.8871	142.8616	165.1714	200.6481
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 200$	μ_R	61.84037	73.53622	94.07807	137.6354
$n = 80$	μ_{ER}	121.8111	142.1197	169.9291	204.592
	$\bar{z}_d$	129.5137	145.6474	170.0913	210.0014
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 300$	μ_R	57.19133	69.61156	92.90621	149.6575
$n = 120$	μ_{ER}	127.9423	156.3544	199.9536	263.6633
	$\bar{z}_d$	138.4083	161.566	200.2256	274.7022

Table 5.5: PRE of z_d family of estimators for Beta

	$X^* \sim B(1,3)$	$Y^* \sim B(3,5)$	$S^* \sim B(6,10)$		
	Estimators	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 50$	μ_R	53.03894	58.2946	64.71001	72.76513
$n = 20$	μ_{ER}	99.43547	109.6728	122.1875	137.8491
	$\bar{z}_d$	112.5182	119.8353	129.4217	142.1943
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 100$	μ_R	52.44889	57.71502	64.1607	72.28161
$n = 40$	μ_{ER}	99.52821	110.0228	122.9181	139.1613
	$\bar{z}_d$	112.9469	120.4867	130.4035	143.6862
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 200$	μ_R	51.23459	56.51788	63.02095	71.27265
$n = 80$	μ_{ER}	99.72638	110.7758	124.5041	142.042
	$\bar{z}_d$	113.8712	121.8982	132.5457	146.9716
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 300$	μ_R	45.86889	55.2673	61.8229	70.20361
$n = 120$	μ_{ER}	100.7399	111.6092	126.2816	145.3237
	$\bar{z}_d$	114.7825	123.4762	134.9644	150.7313

Table 5.6: PRE of z_d family of estimators for Gamma

	$X^* \sim G(6,4)$	$Y^* \sim G(5,3)$	$S^* \sim G(3,10)$		
	Estimators	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 50$	μ_R	82.33888	94.09758	110.2829	134.3089
$n = 20$	μ_{ER}	121.7754	135.5352	152.9642	175.7254
	$\bar{z}_d$	123.3272	135.6996	153.4081	180.3085
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 100$	μ_R	82.00535	93.96859	110.5421	135.3771
$n = 40$	μ_{ER}	122.4635	136.7228	154.9153	178.8997
	$\bar{z}_d$	124.0661	136.8937	155.3801	183.7538
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 200$	μ_R	81.30374	93.69498	111.0984	137.7107
$n = 80$	μ_{ER}	123.9556	139.3239	111.0984	186.0601
	$\bar{z}_d$	125.6702	139.509	159.7551	191.5516
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 300$	μ_R	80.5519	93.3983	111.7117	140.35
$n = 120$	μ_{ER}	125.6254	142.2766	164.2475	194.555
	$\bar{z}_d$	127.4681	142.4785	164.8177	200.8497

Table 5.7: PRE of z_d family of estimators for Poisson

	$X^* \sim P(50)$	$Y^* \sim P(30)$	$S^* \sim P(2)$		
	Estimators	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 50$	μ_R	91.45203	103.3258	121.4688	151.9252
$n = 20$	μ_{ER}	120.6172	134.3155	151.6812	172.6218
	$\bar{z}_d$	121.2778	134.3158	152.907	180.7428
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 100$	μ_R	91.20549	103.3195	121.9548	153.5899
$n = 40$	μ_{ER}	121.2497	135.4093	153.4854	175.4792
	$\bar{z}_d$	121.9313	135.4096	154.767	184.0613
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 200$	μ_R	90.68468	103.3061	94.07807	157.2609
$n = 80$	μ_{ER}	122.6197	137.8000	157.4761	181.8937
	$\bar{z}_d$	123.3473	137.8003	158.8847	191.5541
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 300$	μ_R	90.12327	103.2915	124.1678	161.4711
$n = 120$	μ_{ER}	124.1500	140.5062	162.0732	189.4483
	$\bar{z}_d$	124.93	140.5066	163.6340	200.4562

Table 5.8: PRE of z_d family of estimators for Uniform

	$X^* \sim U(50,100)$	$Y^* \sim U(40,80)$	$S^* \sim U(5,10)$		
	Estimators	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 50$	μ_R	78.22396	89.38644	107.3416	139.1753
$n = 20$	μ_{ER}	106.3031	120.8369	139.8019	162.4823
	$\bar{z}_d$	109.0973	121.6019	139.9128	168.0139
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 100$	μ_R	77.78965	89.0824	107.4855	140.5708
$n = 40$	μ_{ER}	106.8137	121.7922	141.5543	165.4664
	$\bar{z}_d$	109.6794	122.5857	141.6704	171.3305
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 200$	μ_R	76.69485	88.44153	107.7934	143.6392
$n = 80$	μ_{ER}	107.8037	123.8786	145.4381	172.2043
	$\bar{z}_d$	110.8713	124.7355	145.566	178.8521
	μ_0	100.0000	100.0000	100.0000	100.0000
$n' = 300$	μ_R	75.59772	87.75283	108.1314	147.1430
$n = 120$	μ_{ER}	108.9455	126.2377	149.9255	180.2097
	$\bar{z}_d$	112.2261	127.1685	150.0676	187.8486

5.7 Results and Discussion

Table 5.1 provides the parameters of the real populations used to investigate the findings of the theoretical study. Table 5.2 demonstrate the MSE of RRT mean estimator μ_0, RRT ratio estimator given by Sousa *et al.* (2010) μ_R, modified ratio exponential estimator given by Gupta *et al.* (2017) μ_{ER} and the proposed estimator $\bar{z}_d$ along with their respective efficiencies compared to the RRT mean estimator for the two real populations. Under Simulation study, the PRE of existing and proposed estimators for different sample sizes and correlation coefficients are given in Tables 5.3-5.8. For both real and simulated data sets, the suggested class of estimators $\bar{z}_d$ showcased the high value of PRE thus having higher efficiency as compared to the existing estimators in literature.

5.8 Conclusion

In this study, a generalized family of estimators is introduced to estimate the population mean of a sensitive study variable incorporating a non-sensitive supplementary data under two phase sampling design. According to the findings of the simulation study, we observed that being a biased estimator, typical ratio estimators outperform RRT mean estimator particularly for the higher numbers of correlation coefficient whereas for the real data, the ratio estimator consistently outperforms the RRT mean estimator for all given values of correlation coefficient. The findings of the simulation and real data analysis suggest that employing a non-sensitive supplementary information improves the estimation of the population mean of a research variate. Though current estimators such as the normal ratio

RRT estimator outperform the usual RRT mean estimator, the suggested exponential-type estimators provide a substantially higher advantage.

In all instances, irrespective of how strongly the study and supplementary variates are associated, the suggested class of estimators is proven to work more efficiently than other estimators discussed in literature. It is also discovered that if the correlation between the study and the secondary variates is substantial, the efficiency of the suggested family of estimators increases significantly. As a result, the suggested family of estimator can be used to improve the estimation of the population mean of a sensitive variable utilizing known non-sensitive supplementary data in a variety of sensitive areas of research in the Health and Social Sciences. In addition, the suggested family could be extended to include other sampling methodologies.

Chapter 6

Discussions and Conclusions

The study provided the enhanced estimation of the population mean of the study variable that contain sensitive characteristics when an auxiliary variable is present with non-sensitive characteristics.

In Chapter 1, we have provided a breif introduction about RRT which can be used to draw true information on the sensitive issues like illegal activities, trauma and ethically questionable behaviour. We have also discussed about the sampling stategies that have been used in this study. Further, we have also presented different RRT models that have been introduced by researchers over the years. This chapter also includes some fundamental notation, related to the study.

Several authors have worked upon many ratio, regression and exponential type estimators. With this motivation , we have presented a searls type regression estiamtor along with its sampling properties in Chapter 2. We compared the proposed estimator with several other estimators, both theoretically and numerically and proved that the proposed es-

timator contains the highest PRE value, hence proved to be most efficient estimator among the competing estimators under consideration.

Sousa *et al.* (2010) introduced a ratio estimator in sensitive settings. In Chapter 3, we have attempted to generalize Sousa *et al.* (2010) family of estimators by including some well known auxiliary parameters. Further, we have also presented a new improved generalized class of estimators for estimating population mean. We presented the efficiency conditions for the introduced estimators over competing estimators, under which they outperform competing estimators. Two real and simulated data were taken to compare the numerical estimates of PRE of the proposed classes of estimators and we showed that the improved class of estimators performs better than other competing estimators.

As mentioned in Chapter 1, outliers might appear as a result of system behaviour changes, dishonest behaviours, human error or instrument error. A sample might have contained components that came from a population other than the one being studied. Outliers in data can be caused by human mistakes such as those made while collecting, monitoring, or entering data. Authors have developed alternative regression methods like robust estimation in order to address this problem. In Chapter 4, we developed a robust regression type class of estimators for estimating population mean of the sensitive study variable and showed that the robust regression is less susceptible to the effects of outliers and can produce more accurate results in comparison to the conventional OLS. Further, we have proved that the proposed class of estimators performs better than the class of estimators given by Ali *et al.* (2021). Additionally, we have also compared different robust methods namely LTS, LMS, LAD, Tukey M, Hampel M, Huber MM and S method of estimation

numerically.

Taking motivation from Misra and Singh (2016), in Chapter 5, we have proposed a generalized two-phase sampling class of estimators for the improved mean estimation of sensitive variable. We have further provided different members of proposed class of estimators. We compared the theoretical efficiency with RRT ratio, and exponential ratio estimator. Two real and six simulated populations namely normal, log normal, beta, gamma, poisson, and uniform were considered to compared the numerical estimates of PRE. Results has also shown that PRE increases with increase in the correlation between the study and auxiliary variable.

A different sampling method, such as stratified or unequal probability sampling, can be used in future studies for the mean estimation of sensitive variables. In addition, instead of estimating the mean, one can also estimate other parameters such as the variance and distribution function.

Bibliography

Ali, N., Ahma, I., M., Hanif, and Shahzad, U. (2021). Robust-regression-type estimator for improving mean estimation of sensitive variable by using auxiliary information. *Communications in Statistics-Theory and Methods* 50 (4), 979–992.

Anas M. M., Huang Z., Alilah, D.A., Shakqat, A., and Hussain, S. (2021). Mean Estimators Using Robust Quantile Regression and L-Moments Characteristics for Complete and Partial Auxiliary Information. *Mathenatical Problem in Engineering*. DOI: 10.1155/2021/9242895.

Bahl, S. and Tuteja, R. K. (1991). Ratio and Product Type Exponential Estimators. *Journal of Information and Optimization Sciences* 12 (1), 159–164.

Cochran, W. G. (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. *Journal of Agricultural Science* 30, 262–275.

Cochran, W. G. (1942). Sampling theory when the sampling units are of unequal size. *Journal of the American Statistical Association* 37 (218), 199–212.

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

Eichhron, B. H. and Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference* 7 (4), 307–316.

Greenberg, B. G., Kuebler Jr, R. R., Abernathy, J. R., and Horvitz, D.G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association* 66 (334), 243–250.

Grover, L. K. and Kaur, A. (2021). An improved regression type estimator of population mean with two auxiliary variables and its variant using robust regression method. *Journal of Computational and Applied Mathematics* 382, 113072.

Gupta, S. and Shabbir, J. (2004). Sensitivity estimation for personal interview survey questions. *Statistica* 64 (4), 643–653.

Gupta, S., Gupta, B., and Singh, S. (2002). Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and Inference* 100 (2), 239–247.

Gupta, S., Shabbir, J., and Sehra, S. (2010). Mean and sensitivity estimation in optional randomized response model. *Journal of Statistical Planning and Inference* 140 (10), 2870–2874.

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

Gupta, S., Zatezalo, T., and Shabbir, J. (2017). A Generalized mixture estimator of the mean of a sensitive variable in the presence of non-sensitive auxiliary information. *Statistics and Applications* 15 (2), 27–36.

Gupta, S., Zhang, J., Khalil, S., and Sapra, P. (2022). Mitigating lack of trust in quantitative randomized response technique models. *Communications in Statistics-Simulation and Computation*. DOI: 10.1080/03610918.2022.2082477.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics* 42 (6), 1887–1896.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics* 1 (5), 799–821.

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

Kalucha, G., Gupta, S., and 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.

Khoshnevisan, M., Singh, R., Chauhan, P., Sawan, N., and Smarandache, F. (2007). A general family of estimators for estimating population mean using known value of some population parameter(s). *Far East Journal of Theoretical Statistics* 22 (2), 181–191.

Koyuncu, N., S., Gupta, and Sousa, R. (2014). Exponential-type estimators of the mean of a sensitive variable in the presence of non sensitive auxiliary information. *J Communications in Statistics-Simulation and Computation* 43 (7), 1583–1594.

Misra, P. and Singh, R. K. (2016). A generalized double sampling estimator of population mean using variable and attribute both. *India Journal of Applied Research* 6 (1), 344–359.

Murthy, M. N. (1967). *Theory and analysis of sample survey design*. Vol. 16. Calcutta: Statistical Publishing Society.

Onyango, R., Oduor, B., and Francis, O. (2022). Mean Estimation of a Sensitive Variable under Nonresponse Using Three-Stage RRT Model in Stratified Two-Phase Sampling. *Journal of Probability and Statistics*. DOI: 10.1155/2022/4530120.

Oral, E. and Kadilar, C. (2011). Robust ratio-type estimator in simple random sampling. *Journal of Korean Statistical Society* 40, 457–467.

Perri, P. F. (2008). Modified randomized devices for Simmons model. *Model Assisted Statistics and Applications* 3 (3), 233–239.

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

Rousseeuw, P. J. (1984). Least median of square regression. *Journal of the American Statistical Association* 79 (388), 871–880.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection.* Vol. 16. New York: Wiley Series in Probability and Mathematical Statistics.

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

Saleem, I., Sanaullah, A., and Hanif, M. (2019). Double-sampling regression-cum-exponential estimator of the mean of a sensitive variable. *Mathematical Population Studies* 26 (3), 163–182.

Sanaullah, A., Saleem, I., and Shabbir, J. (2019). Use of scrambled response for estimating mean of the sensitivity variable. *Communications in Statistics-Theory and methods* 49 (11), 2634–2647.

Sarndal C. E., Swensson B. and Wretman, J. (1986). *Model assisted survey sampling.* Vol. 16. New York: Springer-Verlag.

Searls, D. T. (1964). The Utilization of a Known Coefficient of Variation in the Estimation Procedure. *Journal of American Statistical Association* 59 (308), 1225–1226.

Sen, A. R. (1978). Estimation of the population mean when the coefficient of variation is known. *Communications in Statistics-Theory and methods* 7 (7), 657–672.

Shahzad, U., Perri, P. F., and Hanif, M. (2019). A new class of ratio-type estimators for improving mean estimation of nonsensitive and sensitive variables by using supplementary information. *Communication in Statistics - Simulation and Computation* 48 (9), 2566–2585.

Shahzad, U., Al-Noor, N. H., Afshan, N., Alilah, D. A., Hanif, M., and Anas, M.M. (2021). Minimum Covariance Determinant-Based Quantile Robust Regression-type Estimators for Mean Parameter. *Mathematical Problem in Engineering*.

Singh, D. and Chaudhary, F. S. (1986). *Sampling theory and methods.* Vol. 16. Amsterdam: New Age International (P) Limited, Publishers.

Singh, G. N., Kumar, A., and Vishwakarma, G. K. (2020a). Estimation of population mean of sensitive quantitative character using blank cards in randomized device. *Communications in Statistics - Simulation and Computation* 49 (6), 1603–1630.

Singh, G. N., Kumar, A., and Vishwakarma, G. K. (2020b). Some alternative additive randomized response models for estimation of population mean of quantitative sensitive variable in the presence of scramble variable. *Communications in Statistics - Simulation and Computation* 49 (11), 2785–2807.

Singh, G.N., Kumar, A., and G.K, Vishwakarma (1998). An alternative estimator for multi-character surveys. An alternative estimator for multi-character surveys. 48 (11), 99–107.

Singh, S. (2003). *Advanced sampling theory with applications. How Michael Selected Amy.* Vol. 16. Amsterdam: Dordrecht: Kluwer Academic Publishers.

Sisodia, B. and Dwivedi, V. (1981). Modified ratio estimator using coefficient of variation of auxiliary Variable. *Journal Indian Society of Agriculture Statistics* 33, 13–18.

Sousa, R., Shabbir, J., Real, P. C., and S., Gupta (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.

Subzar M., Al-Omari A.I. and Alanzi, A. R. A. (2020). The Robust Regression Methods for Estimating of Finite Population Mean Based on SRSWOR in Case of Outliers. *India Journal of Applied Research* 21 (3), 495–510.

Tiwari, N. and Mehta, P. (2017). Additive randomized response model with known sensitivity level. *International Journal of Computational and Theoretical Statistics* 4 (2), 83–93.

Tukey, J. W. (1977). *Exploratory data analysis*. Amsterdam: MA: Addison-Wesley.

Upadhyaya, L. N. and Singh, H. P. (1984). On the estimation of the population mean with known coefficient of variation. *Biometric Journal* 26 (8), 915–922.

Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of American Statistical Association* 79 (309), 63–69.

Waseem, Z., Khan, H., and Shabbir, J. (2021). Generalized exponential type estimator for the mean of sensitive variable in the presence of non-sensitive auxiliary variable. *Communications in Statistics - Theory and Methods* 50 (14), 3477–3488.

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics* 15 (20), 642–656.

Zaman, T. (2019). Improvement of modified ratio estimators using robust regression methods. *Applied Mathematics and Computation* 348, 627–631.

Zaman, T. and Bulut, H. (2019). Modified ratio estimators using robust regression methods. *Communications in Statistics-Theory and methods* 48 (8), 2039–2048.

Zhang, Q., Khalil, S., and Gupta, S. (2021). Mean estimation in the simultaneous presence of measurement errors and non-response using optional RRT models under stratified sampling. *Journal of Statistical Computation and Smiulation* 91 (17), 3492–3504.