SHOU, WENHAO, Ph.D. Kernel Density Estimation Using Randomized Response Models. (2024)
Directed by Dr. Sat Gupta. 102 pp.

The randomized response technique (RRT) was first introduced to estimate prevalence of sensitive characteristics for binary response variables. Extensions to quantitative variables using additive and/or multiplicative scrambling were later explored for population parameter estimation, but estimation of population distribution estimation for sensitive variables remains underexplored.

This dissertation investigates kernel density estimation (KDE) for sensitive variables using additive Randomized Response Technique (RRT) models, addressing the gap in direct distribution estimation in this field. It refines prior work on direct distribution for sensitive variables, particularly KDE under multiplicative RRT models, and explores KDE under additive RRT models. The research encompasses the application of KDE in the presence of auxiliary information and further study of KDE under optional RRT models. Simulation results show that the kernel density estimator using additive scrambling performs better and is more flexible in bandwidth selection compared to multiplicative scrambling. Additionally, the inclusion of auxiliary variables enhances the accuracy of sensitive variable distribution estimation. Introducing sensitivity level W into RRT models as an option proves beneficial under certain conditions for extreme values of W, or when noise levels are high.

By combining the strengths of KDE and additive RRT models, this research seeks to contribute to the advancement of estimation techniques for sensitive variables and provide valuable insights into their distribution. The findings may enhance the understanding and application of survey sampling methodologies when dealing with sensitive and privacy-related information.

KERNEL DENSITY ESTIMATION USING RANDOMIZED RESPONSE MODELS

by

Wenhao Shou

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

Greensboro
2024

Approved by

_____
Committee Chair

I dedicate this dissertation to my beloved parents, whose humor and wit have been a constant source of support and relaxation in every aspect of my life.

APPROVAL PAGE

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

Committee Chair _____

Sat Gupta

Committee Members _____

Sadia Khalil

_____

Sayed Mostafa

_____

Shan Suthaharan

_____

Haimeng Zhang

_____

Date of Acceptance by Committee

_____

Date of Final Oral Examination

ACKNOWLEDGMENTS

This dissertation owes its shape and substance to the invaluable contributions of many individuals.

Firstly, I extend my deepest gratitude to my advisor and mentor, Dr. Sat Gupta. His unwavering enthusiasm for this topic and his encouragement to collaborate have been instrumental. Dr. Gupta has been exceptionally patient and generous with his guidance, suggestions, and time. Throughout every stage of this endeavor, his support and provision of resources enriched my research journey. His insightful and timely feedback kept me focused and productive, even during the most challenging phases.

I also sincerely appreciate my esteemed committee members, Dr. Sadia Khalil, Dr. Sayed Mostafa, Dr. Shan Suthaharan, and Dr. Haimeng Zhang, for their invaluable feedback and guidance. Dr. Khalil's encouragement and wise counsel have been especially impactful, guiding me through various aspects of this dissertation. Dr. Mostafa's unwavering support, particularly during the initial, challenging stages of my research, has been deeply appreciated. Dr. Suthaharan's expertise in computer science significantly streamlined my simulation studies, reducing time costs. Dr. Zhang's support bolstered my confidence in pursuing a Ph.D. in computational mathematics.

Moreover, I would like to acknowledge the invaluable support received from the National Science Foundation (NSF) through the grant DMS-2244160, and from the Department of Mathematics and Statistics for their crucial summer research support and research assistantship. Additionally, I extend my heartfelt gratitude to Monika, Pujita, and Qi for their consistent support during my transition to a new environment, and for their continuous professional guidance, which helped alleviate my apprehensions about teaching. I am also grateful for the warm and supportive environment fostered

by the members of the department, which has contributed significantly to my academic growth.

I also extend my gratitude to all my professors at UNCG, my fellow graduate students, and my friends, both within and outside the program. Special thanks go to my friends Yuanping and Lu for their emotional support and steadfast belief in my success.

Finally, I would like to thank my family for their encouragement and support throughout my academic journey and life.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Sensitive Survey Questions and Related Methodologies

Survey sampling is a crucial tool in social science research, allowing investigators to gather valuable insights into various phenomena by collecting data from a subset of the population. However, the efficacy of this method hinges on the willingness of respondents to provide honest and accurate answers to the survey questions. In the realm of survey sampling, certain questions may tread into sensitive territory, trying to elicit responses that individuals may feel uncomfortable disclosing. This discomfort can stem from various sources, such as concerns about privacy, fear of judgment, or apprehension regarding the potential consequences of divulging certain information (Warner 1965).

The phenomenon of respondents concealing their true responses or opting not to answer altogether, often referred to as non-response or response bias, presents a significant challenge in survey research. Individuals may feel compelled to provide socially desirable responses, aligning their answers with societal norms or expectations rather than expressing their true beliefs or experiences. Additionally, the inherent

nature of sensitive survey questions, which delve into personal or contentious topics, can evoke resistance and reluctance among respondents.

To mitigate the adverse effects of dishonest responses on sensitive survey questions, researchers have developed a repertoire of sampling techniques and methodological approaches. These strategies aim to foster an environment conducive to candid and forthcoming responses while respecting the privacy and comfort of the participants.

In addition to conventional approaches to parameter estimation for sensitive variables, this report delves into the utilization of density estimation techniques, with a particular focus on nonparametric methods such as kernel density estimation (KDE). By adopting a novel perspective, researchers can glean insights into the underlying probability distribution of sensitive variables, thereby enhancing their understanding of respondent behaviors and attitudes. Through a nuanced examination of these methodologies, this dissertation also discusses potential research topics that can be explored in this field.

Over the years, a number of techniques have been proposed and refined to address the inherent challenge of eliciting truthful responses from survey respondents while safeguarding their privacy. This section offers a comprehensive review of several prominent techniques that have been devised to navigate the delicate balance between data integrity and respondent confidentiality.

A seminal contribution to this field was made by Warner (1965), who pioneered the randomized response technique (RRT). This methodological innovation represents a paradigm shift in survey research, offering a nuanced approach to protecting respondent privacy while eliciting honest responses. The RRT functions by introducing controlled noise into respondents' original answers, effectively obfuscating individual responses while still enabling the aggregation of data at a population level. This pre-determined

noise serves as a protective barrier against potential biases stemming from concerns related to social desirability or respondent discomfort, thereby creating an environment conducive to candid and uninhibited information exchange. Since its inception, the RRT has undergone iterative refinements and extensions through the efforts of numerous scholars, a topic that will be examined and contextualized in Chapter 2 of this dissertation.

Furthermore, the evolution of the RRT underscores the dynamic nature of survey methodology, wherein innovative solutions continue to emerge in response to evolving challenges and technological advancements. By elucidating the theoretical underpinnings and practical applications of methodologies such as the RRT, researchers can enhance their understanding of the intricate interplay between data collection, respondent behavior, and ethical considerations. Through rigorous exploration and critical analysis, scholars are poised to contribute to the ongoing refinement and expansion of survey methodologies, thereby fortifying the foundation upon which empirical research is built.

Jones & Sigall (1971) introduced the concept of the bogus pipeline, a sophisticated physiological monitoring device designed to accurately measure the intensity and direction of emotional responses in respondents. Although in reality incapable of such measurements, the device creates an illusion of precision, leading participants to believe that their true emotional states cannot be concealed or manipulated as easily as with traditional paper-and-pencil surveys. Central to the efficacy of the bogus pipeline is the assumption that individuals are disinclined to challenge or deceive a machine, thus fostering a heightened sense of accountability and authenticity in their responses.

By leveraging this technique, researchers can circumvent the limitations of con-

ventional survey methodologies, particularly in contexts where respondents may be inclined to suppress or distort their true emotions. In experimental settings, the implementation of the bogus pipeline has been shown to elicit more candid and uninhibited expressions of negative emotions from participants who would typically exercise restraint or inhibition. This phenomenon underscores the influence of perceived accountability and the desire for authenticity in shaping human behavior within research environments.

In the pursuit of enhancing survey methodology, Raghavarao & Federer (1979) proposed the "black box" (BB) method as an innovative alternative to established techniques such as the randomized response technique (RRT) pioneered by Warner (1965) and its subsequent extensions by scholars like Greenberg et al. (1969), Warner (1971), and Folsom et al. (1973). The BB method introduces a novel approach that capitalizes on supplemented block, (v, k, r, b, λ) balanced incomplete block, and spring balance weighing designs to elicit responses from survey participants in a manner that ensures both data integrity and respondent anonymity.

Central to the BB method is the utilization of block totals, wherein respondents are asked to provide a cumulative count of their responses across a set of pre-determined questions, irrespective of the sensitivity of individual inquiries. Through carefully constructed block designs, encompassing a range of questions and response options, the BB method enables the derivation of estimated responses for each question included in the survey. Importantly, however, this estimation process does not divulge individual responses to specific questions, thereby preserving the anonymity of respondents and safeguarding their privacy.

Given the pressing need to comprehensively explore the impact of social desirability on responses garnered through self-report questionnaires and scales, the development

and dissemination of concise, user-friendly assessment tools hold considerable promise in facilitating researchers' investigations into this pervasive response bias. In this context, Reynolds (1982) advocates for the adoption of the 13-item form, specifically the Marlow-Crowne scale Form C, as a pragmatic solution for assessing social desirability response tendencies within research settings.

The recommendation of the 13-item short form by Reynolds (1982) reflects a deliberate effort to streamline the assessment process without sacrificing the depth or validity of the measurement. By condensing the original scale into a more compact format, researchers are afforded a convenient and efficient means of gauging social desirability biases, thereby expediting data collection and analysis procedures. Moreover, the accessibility and ease of administration associated with the 13-item short form are poised to lower barriers to entry for researchers across disciplines, facilitating broader consideration of social desirability response tendencies in both psychological and sociological research contexts.

The central focus of the research in this dissertation revolves around the utilization of RRT models as indispensable tools for investigating sensitive questions within survey research. RRT models, distinguished by their innovative approach to preserving respondent confidentiality while eliciting candid responses, represent a versatile framework that can be categorized along several dimensions to accommodate varying research objectives and methodological considerations.

Firstly, RRT models can be classified based on the scope of their application in surveys, encompassing three main categories:

- Full RRT: In this approach, all respondents are required to scramble their responses to the survey questions (Warner 1965). By uniformly applying scram-

bling techniques across the entire respondent pool, researchers aim to cultivate an environment conducive to candid and uninhibited disclosure, thereby mitigating concerns related to social desirability bias and respondent reluctance.

- Partial RRT: Unlike the comprehensive implementation of the full RRT, partial RRT involves the selective utilization of randomized response techniques among subsets of respondents within the survey sample MANGAT & SINGH (1990). This nuanced approach allows researchers to tailor the application of RRT methods to specific subpopulations or research contexts where sensitivity levels may vary, optimizing the balance between data integrity and respondent anonymity.

- Optional RRT: Optional RRT provides flexibility for respondents to decide whether to scramble their responses or retain their original answers Gupta et al. (2002). This empowerment of participants acknowledges individual preferences and sensitivities, potentially enhancing the quality and authenticity of responses. Notably, optional RRT has been shown to yield more accurate estimations compared to other RRT variants, as it allows respondents to self-select into the scrambling process based on their comfort levels with the sensitive nature of the questions.

1.2 Density Estimation

Density estimation constitutes a fundamental task in statistical analysis, involving the reconstruction of the underlying probability density function (PDF) from a given set of data points. This endeavor holds significant relevance across various domains, ranging from economics to epidemiology, where understanding the distribution of observed

phenomena is paramount. At its core, density estimation endeavors to uncover the underlying structure of data by discerning patterns and trends inherent within the sample.

1.2.1 Parametric Density Estimation

Parametric methods represent a powerful approach to density estimation, founded on the assumption that the underlying population distribution conforms to a specific parametric form characterized by a finite set of parameters. These methodologies, such as the method of moments (MOM) by Pearson (1894) and maximum likelihood estimation (MLE), form the cornerstone of statistical inference by facilitating the estimation of distribution parameters from observed data.

The method of moments (MOM) operates on the principle of equating population moments, which are mathematical functions of the distribution parameters, with their corresponding sample moments derived from the observed data. By aligning these moments, MOM enables the estimation of distribution parameters through the Law of Large Numbers, leveraging the convergence of sample moments to their population counterparts as the sample size increases. This approach offers a straightforward and intuitive means of parameter estimation, particularly in cases where analytical solutions are readily attainable.

Conversely, maximum likelihood estimation (MLE) represents a more sophisticated framework for parameter estimation, predicated on the principle of maximizing the likelihood function given the observed sample data. In essence, MLE seeks to identify the parameter values that render the observed data most probable under the assumed parametric model. By optimizing the likelihood function, MLE provides estimates

that are asymptotically unbiased, efficient, and consistent, making it a versatile and widely-used tool in statistical modeling and inference.

The distinguishing feature of MLE lies in its emphasis on the likelihood of observing the sample data, rather than the moments of the underlying distribution. This affords MLE greater flexibility in accommodating complex data structures and distributional forms, while also enabling the incorporation of prior information through the specification of appropriate likelihood functions. Despite its computational intensity and reliance on iterative optimization techniques, MLE offers a robust and versatile framework for parameter estimation across a wide range of statistical models and applications.

In summary, parametric methods such as MOM and MLE offer powerful tools for density estimation by leveraging the assumed parametric form of the underlying distribution. While MOM provides a straightforward approach based on moment matching, MLE offers a more flexible and sophisticated framework grounded in likelihood maximization. By harnessing the principles of statistical inference, these methodologies enable researchers to extract valuable insights from observed data and make informed decisions in various scientific and practical contexts.

1.2.2 Non-parametric Density Estimation

While parametric methods offer a streamlined approach to density estimation by leveraging assumptions about the population distribution, non-parametric methods adopt a more flexible stance by eschewing such assumptions altogether. Instead of constraining the analysis to a predefined distributional form, non-parametric methods endeavor to estimate the density across the entire sample space, thereby accommodating diverse

data structures and distributions. Kernel density estimation (KDE) proposed by Rosenblatt (1956) and Parzen (1962) emerges as a prominent non-parametric technique, renowned for its versatility and efficacy in capturing the underlying structure of data.

KDE operates by assigning a kernel function to each data point, with the density estimate at any given point being determined by the weighted contributions of neighboring data points. The kernel density estimate is given by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x - X_i}{h}), \tag{1.1}$$

where the kernel function $K(x)$, typically symmetric and bounded, governs the spread of influence exerted by each data point, while the bandwidth parameter h regulates the smoothness of the resulting density estimate. As presented in (1.1), the KDE estimate at a specific point is computed as the weighted average of kernel functions centered around each data point, normalized by the sample size and bandwidth.

Importantly, KDE exhibits a number of advantages over parametric approaches, particularly in scenarios characterized by small sample sizes or non-standard data distributions. By leveraging the entirety of the sample data, KDE offers a robust and flexible means of estimating the underlying density function, thereby facilitating nuanced insights into the distributional characteristics of the observed phenomena. Furthermore, as the sample size increases, the KDE approximation tends to converge towards the true PDF under certain conditions, underscoring its efficacy in capturing the underlying structure of data with increasing fidelity.

1.3 Outline of the Dissertation

In this dissertation, the central objective revolves around advancing the methodology of estimating the entire distribution of sensitive variables through KDE within the framework of RRT models. This approach aims to provide a comprehensive understanding of the underlying distribution, offering broader insights compared to traditional point estimators.

Chapter 2 will offer a comprehensive literature review, consolidating current research on RRT models, KDE, and their convergence within the realm of sensitive survey research.

In Chapters 3 - 5, the research is organized into several key areas of investigation, each delving into distinct aspects of KDE under various RRT models:

- Validation and Expansion of Theoretical Findings: Chapter 3 of the dissertation will be dedicated to validating and expanding upon the theoretical foundations laid by prior research concerning KDE under multiplicative RRT models. Building upon existing literature, this Chapter seeks to validate the theoretical framework of KDE in the context of multiplicative RRT, while exploring new avenues for improvement and extension.

  In continuation, Chapter 3 will also extend its focus to the application of KDE under more commonly-used additive RRT models, which are widely regarded as superior to multiplicative RRT models in certain contexts such as privacy protection and ease of implementation.

- Incorporation of Auxiliary Information: Chapter 4 will delve into the integration of auxiliary information into additive RRT models when utilizing KDE. By

leveraging auxiliary data sources, this Chapter aims to enhance the accuracy and efficiency of KDE-based estimators.

- Exploration of KDE under Optional RRT Models: Chapter 5 will explore the application of KDE within the context of optional RRT models. This Chapter will highlight the advantages of optionality, which generally leads to better estimation outcomes compared to non-optionality in most cases.

Chapter 6 will offer a general discussion of the research introduced in this dissertation. It will also summarize the most significant findings and some future directions for the work presented in this dissertation.

Chapter 2: Literature Review

As discussed in Section 1.1, the randomized response technique (RRT) was initially introduced by Warner (1965) as a strategic response to mitigate the influence of social desirability bias (SDB), particularly concerning binary sensitive variables. Over time, this foundational RRT framework has undergone extensive refinement and expansion by numerous researchers, evolving to accommodate the study of quantitative sensitive variables through the incorporation of additive and/or multiplicative scrambling mechanisms. These include Warner (1971), Pollock & Bek (1976), Eichhorn & Hayre (1983), Gupta et al. (2002), Diana & Perri (2011), Blair et al. (2015), Gupta et al. (2018), and Khalil et al. (2021). A considerable body of research has been dedicated to estimating population parameters, such as the mean or variance, of distributions governed by RRT models, by researchers including Gupta et al. (2010), Sousa et al. (2010), Gupta et al. (2012), Khalil, Gupta & Hanif (2018), Mehta & Aggarwal (2018), Gupta et al. (2020), and Narjis & Shabbir (2020). However, there remains a notable gap in the literature concerning the direct estimation of the underlying distribution of these sensitive variables.

In this Chapter, we aim to provide a comprehensive overview of prior research endeavors focusing on the estimation of population parameters, with a primary emphasis on mean estimators, within the framework of various RRT models. These studies have

played a pivotal role in elucidating the nuances of parameter estimation under RRT methodologies, offering valuable insights into the challenges and opportunities inherent in this domain. Additionally, we will synthesize the limited existing literature on the estimation of density curves under RRT models, highlighting notable contributions and potential avenues for further exploration.

By synthesizing and contextualizing prior research findings, this section seeks to lay the groundwork for the subsequent discussion on the estimation of sensitive variable distributions within RRT frameworks. Through a systematic examination of existing methodologies and empirical findings, we aim to elucidate the current landscape of knowledge in this field by pinpointing existing gaps and opportunities for future research initiatives. Ultimately, this endeavor seeks to enhance our comprehension of statistical inference within RRT models, thus fostering more resilient and well-informed decision-making within the sphere of sensitive survey research.

2.1 Estimating Population Parameters of a Sensitive Variable

2.1.1 Binary RRT Models

Warner's Binary Model

In his seminal work, Warner (1965) introduced the Randomized Response Technique (RRT) as a groundbreaking method aimed at enhancing response rates in surveys addressing sensitive topics. Recognizing the inherent challenges associated with eliciting truthful responses on sensitive matters, Warner devised a novel approach to

mitigate response bias while safeguarding respondent privacy. The essence of Warner's innovation lies in the utilization of randomization devices, which afford respondents a degree of anonymity and confidentiality in their responses.

At the core of Warner's binary model is the strategic implementation of random prompts, whereby respondents are randomly directed to answer sensitive questions either directly or indirectly (see Figure 2.1). This randomized prompting mechanism serves to obscure the true nature of the questions being answered, thereby preserving respondent privacy and confidentiality. By randomizing the prompt process, the investigator remains unaware of the specific question posed to each respondent, ensuring that only the final response (yes/no) is known.

Central to Warner's design is the pre-determined proportion of respondents prompted to answer the direct or the indirect question, a crucial element in unraveling the overall survey results. By carefully controlling the ratio of respondents subjected to each prompting condition, Warner effectively balances the trade-off between privacy protection and data accuracy. This pre-determined proportion serves as a key parameter in the estimation process, enabling researchers to derive unbiased estimates of the prevalence of sensitive traits within the population.

Let

$\pi =$ *Proportion of respondents with the sensitive trait*

$p_y =$ *Probability of a "yes" response*

$p =$ *Probability of respondent answering the sensitive question directly,* $p \neq \frac{1}{2}$

$1 - p =$ *Probability of respondent answering the sensitive question indirectly*

$n =$ *Sample size*

Then

$$p_y = p\pi + (1 - p)(1 - \pi). \tag{2.1}$$

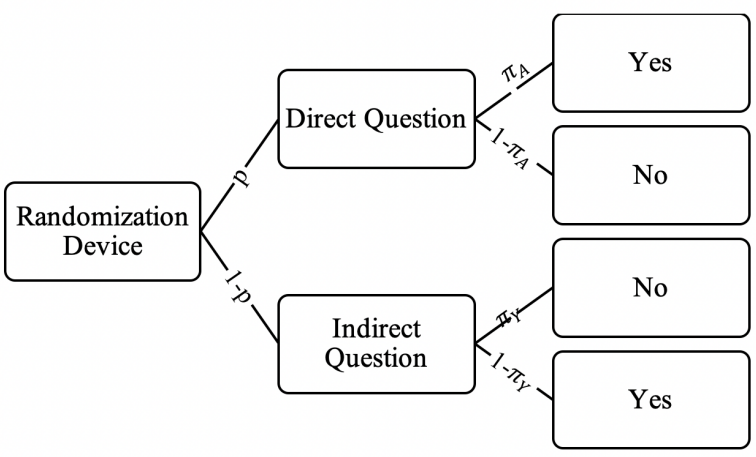

Figure 2.1. Warner's Binary RRT Model

This leads to an estimate of π (defined for $p \neq \frac{1}{2}$) which is given as

$$\hat{\pi} = \frac{\hat{p_y} - (1-p)}{2p-1}. \tag{2.2}$$

The variance of the above estimator under simple random sampling with replacement (SRSWR) is

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}. \tag{2.3}$$

In essence, Warner's pioneering work on RRT revolutionized survey methodology by introducing a nuanced approach to addressing response bias and safeguarding respondent privacy. By harnessing the power of randomization and pre-determined proportions, Warner's binary model offers a robust framework for eliciting truthful responses on sensitive topics, thereby advancing the integrity and reliability of survey research in sensitive domains.

Binary Unrelated-Question Model

Expanding upon the groundwork laid by Warner (1965), Greenberg et al. (1969) introduced an innovative extension to the unrelated-question design within the framework of RRT. Departing from the conventional approach of employing two relevant questions to partition the sample, Greenberg proposed a novel method that divided the sample into two distinct groups differently.

In Greenberg's design (Figure 2.2), respondents were presented with two options, one of which pertained to a non-sensitive, innocuous attribute unrelated to the sensitive topic under investigation. By introducing this unrelated attribute, Greenberg sought to bolster respondents' confidence in the anonymity afforded by the technique, thereby fostering a greater willingness to provide truthful responses. The rationale behind this approach lies in the notion that by coupling the sensitive question with an innocuous alternative, respondents perceive a heightened level of privacy protection, which in turn encourages more honest and forthcoming responses.

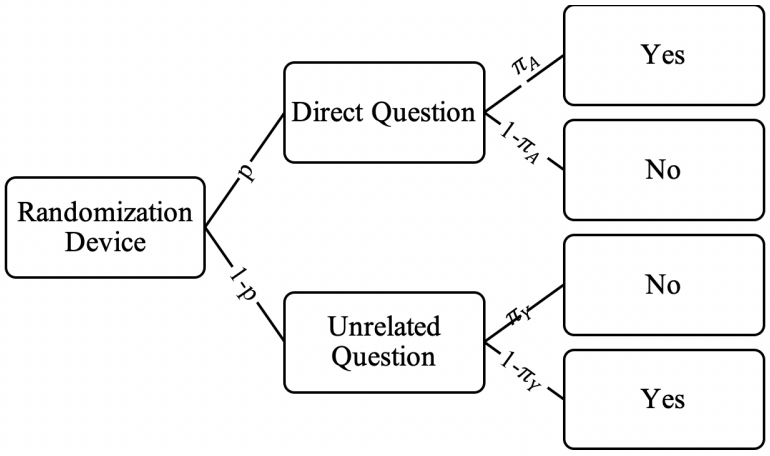

Figure 2.2. Binary Unrelated-Question Model

Let

π_A = *Proportion of respondents with the sensitive trait A*

π_Y = *Proportion of respondents with the unrelated trait Y*

λ = *Probability of a "yes" response*

p = *Probability of respondent answering the sensitive question directly*

$1 - p$ = *Probability of respondent answering the unrelated question*

Then

$$\lambda = p\pi_A + (1 - p)(1 - \pi_Y). \tag{2.4}$$

This leads to an estimate of π_A which is given as

$$\hat{\pi_A} = \frac{\hat{\lambda} - (1 - p)\pi_Y}{p}. \tag{2.5}$$

with the variance, under SRSWR, given by

$$Var(\hat{\pi_A}) = \frac{\lambda(1 - \lambda)}{np^2}. \tag{2.6}$$

This unrelated-question design represents a strategic refinement aimed at enhancing the efficacy of the RRT in eliciting accurate and reliable data on sensitive topics. By leveraging the psychological principle of perceived anonymity, Greenberg's approach capitalizes on respondents' perceptions of privacy and confidentiality to engender greater disclosure and candor in their responses.

In essence, Greenberg's contribution underscores the importance of psychological factors in shaping respondent behavior within the context of survey research. By incorporating elements that bolster respondents' confidence in the confidentiality of their responses, Greenberg's extension to the unrelated-question design represents a

significant advancement in the methodology of RRT, offering a nuanced approach to mitigating response bias and enhancing the veracity of survey data.

Binary Unrelated-Question Model With Untruthfulness

Both of the RRT models discussed above operate under the assumption of complete respondent honesty. However, in real-world scenarios, this assumption may not always hold true, leading to potential biases in the estimation process. Situations where respondents may be inclined to provide untruthful responses can arise, particularly in the context of highly sensitive questions or when adequate respondent training is lacking.

To address the challenge of untruthful responses, Young et al. (2019) proposed a binary uncorrelated-question model as an additional safeguard within the RRT framework. This innovative model introduces an extra layer of precaution to mitigate the impact of respondent dishonesty on the estimation of sensitive traits.

The binary uncorrelated-question model, as developed by Young et al. (2019), entails a two-question design aimed at discerning and redirecting potentially untruthful responses. In the first question (Figure 2.3), respondents are queried about their trust in the randomization process, drawing upon the methodology outlined by Greenberg et al. (1969). This initial inquiry serves as a precursor to gauge respondents' willingness to engage honestly in the survey process.

Subsequently, in the second question (Figure 2.4), respondents are presented with a choice: if they express trust in the randomization process, they are directed to respond using the established Greenberg et al. (1969) model for sensitive questions. Conversely, if respondents indicate a lack of trust in the randomization process, they are redirected to answer an unrelated question, thereby circumventing the sensitive

inquiry altogether.

This two-question design effectively leverages respondent trust as a determinant for the appropriate course of action in the survey process. By offering respondents an alternative response pathway in cases of mistrust or potential dishonesty, the binary uncorrelated-question model introduces a strategic mechanism for safeguarding the integrity of survey data.

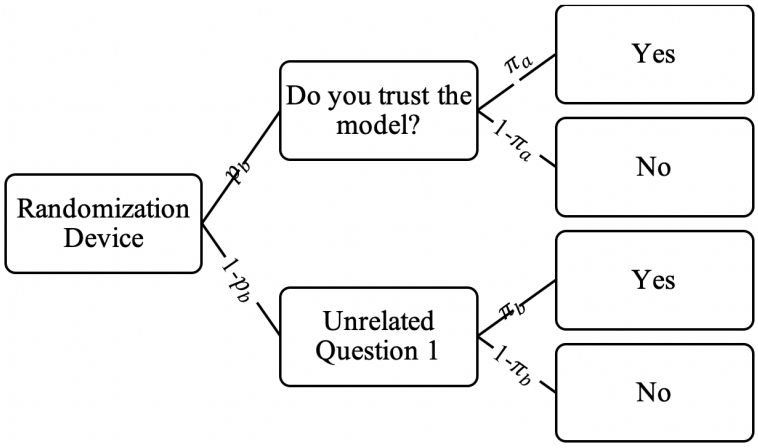

Figure 2.3. Binary Unrelated-Question Model With Untruthfulness (Question 1)

Let

π_a = *Proportion of respondents who trust the model*

π_x = *Proportion of respondents with the sensitive trait*

π_y = *Proportion of respondents with some unrelated trait*

π_b = *Proportion of respondents with some other unrelated trait*

p_{yi} = *Probability of a "yes" response to Question* i $(i = 1, 2)$

p_b = *Probability of respondent answering the question about trust in Question 1*

p = *Probability of respondent answering the sensitive question directly in Question 2*

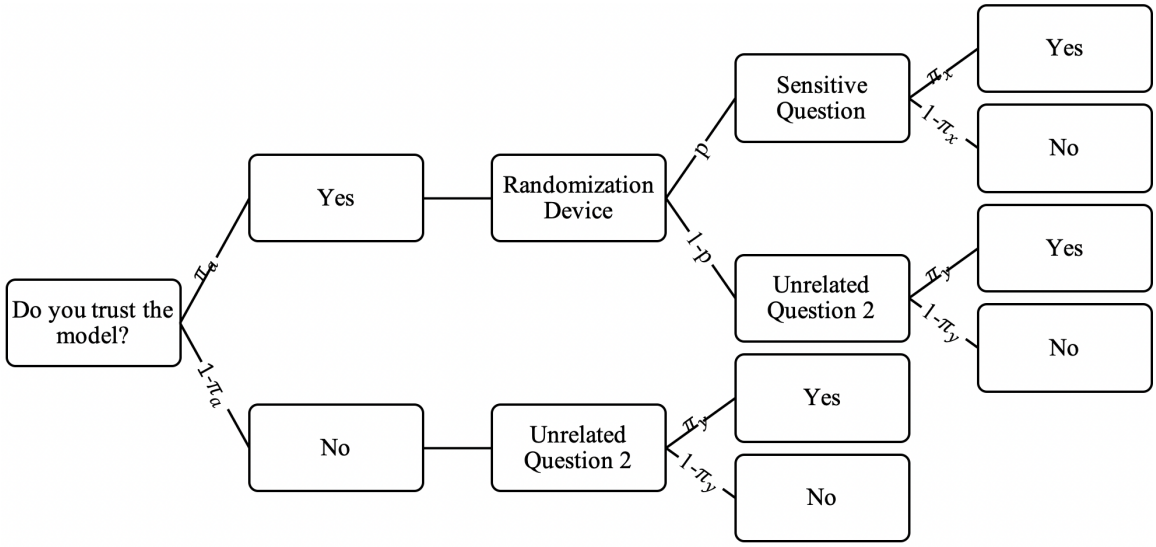

Figure 2.4. Binary Unrelated-Question Model With Untruthfulness (Question 2)

Then

$$p_{y1} = p_b\pi_a + (1 - p_b)\pi_b, \tag{2.7}$$

$$p_{y2} = \pi_a[p\pi_x + (1 - p)\pi_y] + (1 - \pi_a)\pi_y. \tag{2.8}$$

This leads to

$$\hat{\pi_a} = \frac{\hat{p_{y1}} - (1 - p_b)\pi_b}{p_b} \quad \text{and} \quad \hat{\pi_x} = \frac{\hat{p_{y2}} - \pi_y(1 - \hat{\pi_a}p)}{\hat{\pi_a}p} \tag{2.9}$$

with the variances given by

$$Var(\hat{\pi_a}) = \frac{p_{y1}(1 - p_{y1})}{np_b} \quad \text{and} \quad Var(\hat{\pi_x}^*) = \frac{p_{y2}(1 - p_{y2})}{n(\pi_a p)^2} + \frac{p_{y1}(1 - p_{y1})p^2(\pi_y - p_{y2})^2}{np_b^2(\pi_a p)^4}, \tag{2.10}$$

where $\hat{\pi_x}^*$ is an unbiased estimator of π_x up to first-order approximation.

In essence, Young et al. (2019)'s innovative approach represents a proactive measure to address the complexities inherent in survey research, particularly in contexts where

respondent honesty may be compromised. Through the integration of a binary uncorrelated-question model within the RRT framework, researchers gain a valuable tool for mitigating the impact of respondent bias on the estimation of sensitive traits, thereby enhancing the reliability and validity of survey findings.

2.1.2 Quantitative RRT Models

In addition to binary RRT models, there exists another significant category known as quantitative RRT models, based on the type of questions posed and the potential responses elicited. Unlike binary models that only accommodate yes/no responses, quantitative RRT models expand the spectrum by allowing respondents to provide numerical values as their answers. This broader range of potential responses introduces a greater degree of flexibility in the scrambling process, as the responses can vary along a continuous scale rather than being limited to a binary choice.

The exploration of quantitative RRT models has received considerable attention in the research community over the past decades, with numerous studies dedicated to understanding their intricacies and applications. These studies delve into various aspects of quantitative RRT models, shedding light on their theoretical foundations, practical implementation, and implications for survey research methodologies. Below, we discuss some noteworthy studies that have contributed to the advancement of quantitative RRT models.

Greenberg's Quantitative Model

This quantitative model introduced by Greenberg et al. (1971) builds upon the foundation laid by the unrelated-question model pioneered by Greenberg et al. (1969)

for binary data. In essence, it represents an extension and refinement of the earlier model, catering to the nuanced complexities inherent in survey research on sensitive topics.

In the framework proposed by Greenberg et al. (1971), a pre-determined fraction of respondents is tasked with answering the sensitive question, while the remaining respondents are presented with an unrelated query. This strategic allocation of questions serves to diversify the survey instrument, mitigating potential biases and enhancing the reliability of responses.

The rationale behind this approach lies in its capacity to create a balanced survey design that safeguards respondent privacy while eliciting meaningful data. By incorporating both sensitive and unrelated questions, the model seeks to minimize the perceived intrusiveness of the survey, thereby fostering a more conducive environment for candid responses.

Let

$Y = Response\ to\ the\ sensitive\ question$

$B = Response\ to\ the\ unrelated\ question$

$p = Probability\ of\ respondent\ answering\ the\ sensitive\ question$

$1 - p = Probability\ of\ respondent\ answering\ the\ unrelated\ question$

$\mu_Y = Population\ mean\ of\ the\ sensitive\ variable$

$\mu_B = Population\ mean\ of\ the\ unrelated\ variable$

$n = Sample\ size$

$Z = Reported\ response,\ where$

$$Z = \begin{cases} Y & with\ probability\ p \\ B & with\ probability\ 1-p \end{cases} \tag{2.11}$$

Then

$$\bar{Z} = E(Z) = p\mu_Y + (1-p)\mu_B. \tag{2.12}$$

This leads to

$$\hat{\mu_{Y,Greenberg}} = \frac{\bar{Z} - (1-p)\mu_B}{p}. \tag{2.13}$$

with the variance, under SRSWR, given by

$$Var(\hat{\mu_{Y,Greenberg}}) = \frac{\sigma_Z^2}{np^2}. \tag{2.14}$$

Warner's Quantitative Model

In contrast to the quantitative model proposed by Greenberg et al. (1971), which extends the binary model, Warner (1971) introduced a novel approach that involves incorporating a random number drawn from a known distribution to the sensitive value. This pioneering concept marked a significant departure from conventional survey methodologies, offering a fresh perspective on addressing response bias in sensitive surveys.

The innovation put forth by Warner (1971) received considerable attention within the research community, prompting further exploration and refinement of the additive RRT model. Building upon Warner's foundational work, Pollock & Bek (1976) delved deeper into the additive RRT model, elucidating its underlying principles and exploring its implications for survey research. While Pollock's focus primarily centered on the additive RRT model, brief mention was made of the multiplicative RRT models, hinting at the broader landscape of RRT methodologies.

Expanding upon the groundwork laid by Warner (1971) and Pollock & Bek (1976), Eichhorn & Hayre (1983) continued the investigation into the multiplicative RRT model,

further advancing our understanding of its intricacies and applications. Eichhorn's research contributed valuable insights into the theoretical underpinnings and practical considerations associated with multiplicative RRT models, shedding light on their potential utility in sensitive survey research.

Collectively, the contributions of Warner (1971), Pollock & Bek (1976), and Eichhorn & Hayre (1983) have moved the development and refinement of RRT methodologies further, enriching the methodological toolkit available to researchers in sensitive survey domains. By exploring alternative approaches to question design and response elicitation, these scholars have paved the way for more nuanced and effective strategies for collecting data on sensitive topics.

Additive Model

Pollock & Bek (1976) further developed the concept of additive scrambling within quantitative RRT models. In the additive model proposed by Pollock & Bek (1976), respondents are tasked with summing their response to a sensitive question with a randomly generated value drawn from a known distribution.

Let

$Y =$ *Response to the sensitive question*

$S =$ *Scrambling variable (independent of* Y*)*

$\mu_Y =$ *Population mean of the sensitive variable*

$\mu_S =$ *Population mean of the scrambling variable*

$\sigma_Y^2 =$ *Variance of the sensitive variable*

$\sigma_S^2 =$ *Variance of the scrambling variable*

$Z =$ *Reported response, where* $Z = Y + S$

Then

$$E(Z) = E(Y) + E(S) = \mu_Y + \mu_S. \tag{2.15}$$

An unbiased estimator of mean of sensitive variable is given by

$$\mu_{Y,\hat{W}arner} = \bar{Z} - \mu_S. \tag{2.16}$$

The variance of this estimator, under SRSWR, is given by

$$Var(\mu_{Y,\hat{W}arner}) = Var(\bar{Z}) = \frac{\sigma_Z^2}{n} = \frac{\sigma_Y^2}{n} + \frac{\sigma_S^2}{n}. \tag{2.17}$$

By incorporating random additive components into respondents' answers, this model introduces a level of noise that obscures the true nature of the sensitive question, thus safeguarding respondent privacy while preserving the integrity of the survey data.

Multiplicative Model

Eichhorn & Hayre (1983) carried forward the investigations initiated by Pollock & Bek (1976), building upon their groundwork to introduce a novel quantitative RRT model centered on multiplicative scrambling. In this innovative model, respondents are instructed to multiply their true response to a sensitive question by a randomly generated number drawn from a pre-defined distribution. This multiplication process injects a degree of variability into the respondent's answer, thereby enhancing the confidentiality and privacy of their responses.

Let

$Y =$ *Response to the sensitive question*

$S =$ *Scrambling variable (independent of* X*)*

μ_Y = *Population mean of the sensitive variable*

θ = *Population mean of the scrambling variable*

σ_Y^2 = *Variance of the sensitive variable*

γ^2 = *Variance of the scrambling variable*

Z = *Reported response, where* $Z = YS$

An unbiased estimator of mean of sensitive variable is given by

$$\hat{\mu_{Y,multi}} = \frac{\bar{Z}}{\theta} \tag{2.18}$$

with the variance, under SRSWR, given by

$$Var(\hat{\mu}_{Y,multi}) = \frac{\sigma^2 + (\gamma/\theta)^2 E(Y^2)}{n}. \tag{2.19}$$

Eichhorn's exploration of the multiplicative RRT model represents a notable advancement in survey methodology, offering a refined approach to mitigating response bias in sensitive surveys. By incorporating multiplicative scrambling techniques, this model introduces a layer of randomness that obfuscates the individual's true answer, thereby safeguarding respondent privacy while preserving the integrity of the survey data.

Optional Multiplicative Model

In the RRT models explored thus far, there exists an implicit assumption that all respondents perceive the posed question as sensitive. However, the reality may differ, as a subset of respondents might not perceive the question as sensitive and may be inclined to provide their true responses without the need for scrambling. This nuanced

scenario prompted Gupta et al. (2002) to introduce the concept of optionality into the multiplicative RRT model proposed by Eichhorn & Hayre (1983).

Within this innovative model, respondents are afforded the flexibility to choose between providing a true response or opting for a scrambled response to certain questions. This introduction of choice grants respondents agency over their responses, allowing them to exercise discretion based on their individual perceptions of question sensitivity. Importantly, the investigator remains unaware of the type of response provided by each respondent, preserving respondent confidentiality and ensuring the integrity of the survey data.

Let

$Y =$ *Response to the sensitive question*

$S =$ *Scrambling variable*

$\mu_Y =$ *Population mean of the sensitive variable*

$\mu_S =$ *Population mean of the scrambling variable*

$\sigma_S^2 =$ *Variance of the scrambling variable*

$W =$ *Sensitivity level (Proportion of people who find the question sensitive)*

$Z =$ *Reported response, where*

$$Z = \begin{cases} YS & with\ probability\ W \\ Y & with\ probability\ 1-W \end{cases} \tag{2.20}$$

Let $\mu_S = 1$, then $E(Z) = \mu_Y$, and one can obtain an estimate of μ_Y, given by

$$\hat{\mu_Y} = \bar{Z} \tag{2.21}$$

with the variance, under SRSWR, given by

$$Var(\hat{\mu_Y}) = \frac{\sigma_Y^2 + W\sigma_S^2(\sigma_Y^2 + \mu_Y^2)}{n}. \tag{2.22}$$

An estimator of W is given by

$$\hat{W} = \frac{\frac{1}{n}\Sigma_{i=1}^{n} log(Z_i) - log(\frac{1}{n}\Sigma_{i=1}^{n} Z_i)}{E[log(S)]}. \tag{2.23}$$

The incorporation of optionality into the multiplicative RRT model represents a significant advancement in survey methodology, offering a refined approach to addressing variability in respondent perceptions of question sensitivity. By acknowledging and accommodating individual differences in sensitivity perception, this model enhances the accuracy and reliability of the survey results, thereby enriching the overall quality of data collected.

Furthermore, the optional model enables the estimation of the sensitivity level associated with each question, providing valuable insights into the distribution of sensitivity perceptions within the surveyed population. This estimation process contributes to a deeper understanding of respondent behavior and attitudes, facilitating more nuanced analyses and interpretations of survey findings.

In essence, the introduction of optionality into the multiplicative RRT model reflects a commitment to methodological innovation and responsiveness to the complexities of real-world survey settings. By empowering respondents with choice and discretion, this model promotes transparency, trust, and cooperation, fostering a more conducive environment for candid responses and meaningful data collection.

Two-Stage Optional Additive Model

Considering that a non-zero scrambled response within multiplicative RRT models often signifies the existence of sensitive behavior to some degree, additive RRT models emerge as a preferable choice in numerous instances. The rationale behind this preference lies in the enhanced privacy protection afforded by additive models, which effectively obscure the true nature of respondents' sensitive behaviors while ensuring the integrity of the survey data. Moreover, additive models offer a more user-friendly approach, particularly for survey participants with limited mathematical proficiency, as they entail simpler computational procedures and intuitive response mechanisms.

In recognition of the practical advantages offered by additive RRT models, Gupta et al. (2010) proposed an optional additive RRT model to address scenarios where respondents may benefit from additional privacy safeguards or encounter challenges in navigating the survey process due to mathematical constraints. By offering respondents the flexibility to opt for either their true responses or scrambled responses, the model empowers individuals to exercise informed choices regarding their participation in sensitive surveys.

Furthermore, the optional additive RRT model employs a split-sample methodology to concurrently estimate the sensitivity level without approximations. This approach ensures robust and precise estimation, thereby enhancing the reliability and accuracy of survey findings. By integrating advanced statistical techniques with user-centric design principles, Gupta's model represents a significant advancement in survey methodology, offering researchers a comprehensive toolkit for navigating the intricacies of sensitive data collection with integrity and confidence.

Let

Y = *Response to the sensitive question*

S_i = *Scrambling variable used for the* i^{th} *sub-sample* (i = 1, 2)

μ_Y = *Population mean of the sensitive variable*

θ_i = *Population mean of the scrambling variable* S_i (i = 1, 2, $\theta_1 \neq \theta_2$)

$\sigma^2_{S_i}$ = *Variance of the scrambling variable* S_i (i = 1, 2)

n_i = *Size of the* i^{th} *sub-sample* (i = 1, 2)

W = *Sensitivity level (Proportion of people who find the question sensitive)*

Z = *Reported response, where*

$$Z_i = \begin{cases} Y + S_i & with\ probability\ W \\ Y & with\ probability\ 1 - W \end{cases}, i = 1, 2 \tag{2.24}$$

Then the mean for Z_i is given by

$$E(Z_i) = \mu_Y + \theta_i W, \tag{2.25}$$

where $E(S_i) = \theta_i$.

This leads to

$$\hat{\mu_Y} = \frac{\theta_1 \bar{Z}_2 - \theta_2 \bar{Z}_1}{\theta_1 - \theta_2} \tag{2.26}$$

and

$$\hat{W} = \frac{\bar{Z}_1 - \bar{Z}_2}{\theta_1 - \theta_2}. \tag{2.27}$$

with the variances, under SRSWR, given by

$$Var(\hat{\mu_Y}) = \frac{1}{(\theta_2 - \theta_1)^2}\left[\theta_2^2\frac{\sigma_{Z_1}^2}{n_1} + \theta_1^2\frac{\sigma_{Z_2}^2}{n_2}\right] \tag{2.28}$$

and

$$Var(\hat{W}) = \frac{1}{(\theta_2 - \theta_1)^2}\left[\frac{\sigma_{Z_1}^2}{n_1} + \frac{\sigma_{Z_2}^2}{n_2}\right], \tag{2.29}$$

where $\sigma_{Z_i}^2 = \sigma_Y^2 + \sigma_{S_i}^2 W + \theta_i^2 W(1-W)$.

Linear Combination Model

Diana & Perri (2011) undertook a comprehensive review and comparative analysis of the diverse quantitative RRT models documented in existing literature. Drawing upon a wide array of research findings and methodologies, their review sought to elucidate the strengths, weaknesses, and unique characteristics of each model, thereby providing valuable insights into the landscape of quantitative RRT methodologies.

Building upon their extensive review, Diana & Perri (2011) proposed a unified framework for conducting surveys on sensitive topics with quantitative responses. This general model represents a synthesis of the most effective strategies and techniques identified through their comparative analysis, offering researchers a versatile and adaptable approach to addressing response bias and safeguarding respondent privacy.

A notable feature of the model proposed by Diana & Perri (2011) is the integration of both additive and multiplicative scrambling techniques. By combining these two approaches, the model aims to optimize the balance between privacy protection and survey efficiency. Additive scrambling introduces variability into respondents' answers by adding a random value to their true response, while multiplicative scrambling scales the response by a random factor, further obscuring the true nature of the sensitive

question.

The rationale behind this combined approach lies in its capacity to offer a comprehensive and robust solution to the challenges inherent in survey research on sensitive topics. By leveraging the complementary strengths of additive and multiplicative scrambling, the proposed model seeks to maximize respondent confidentiality while minimizing the impact on data accuracy and survey efficiency.

Let

$Y =$ *Response to the sensitive question*

$S =$ *Additive scrambling variable*

$D =$ *Multiplicative scrambling variable*

$\mu_Y =$ *Population mean of the sensitive variable*

$\mu_S =$ *Population mean of the scrambling variable* S

$\mu_D =$ *Population mean of the scrambling variable* D

$\sigma_Y^2 =$ *Variance of the sensitive variable*

$\sigma_S^2 =$ *Variance of the scrambling variable* S

$\sigma_D^2 =$ *Variance of the scrambling variable* D

$Z =$ *Reported response, where* $Z = DY + S$

Also, Y, S, and D are mutually independent.

Let $\mu_S = 0$ and $\mu_D = 1$, then $E(Z) = \mu_Y$, and one can obtain an estimate of μ_Y, given by

$$\hat{\mu_Y} = \bar{Z} \tag{2.30}$$

with the variance, under SRSWR, given by

$$Var(\hat{\mu_Y}) = \frac{\sigma_D^2(\mu_Y^2 + \sigma_Y^2) + \sigma_Y^2 + \sigma_S^2}{n}. \tag{2.31}$$

In essence, the work of Diana & Perri (2011) represents a significant contribution to the field of survey methodology, offering researchers a practical and effective framework for conducting surveys on sensitive topics with quantitative responses. Through their review, comparative analysis, and proposal of a general model, Diana & Perri (2011) have advanced our understanding of quantitative RRT methodologies and provided valuable guidance for researchers seeking to navigate the complexities of survey research in sensitive domains.

2.1.3 Use of Auxiliary Information Under Quantitative RRT Models

In survey sampling, auxiliary information serves as a valuable tool for refining the accuracy of estimators used to infer population parameters from finite samples. When auxiliary variables exhibit a strong correlation with the variable of interest, leveraging this additional information can lead to more precise estimates of population parameters. Consequently, in scenarios where data on a non-sensitive variable correlates closely with a sensitive variable, integrating auxiliary information into quantitative RRT models can yield improved estimations.

Numerous studies have explored the integration of auxiliary information into quantitative RRT models, seeking to enhance the robustness and accuracy of survey estimates. By harnessing auxiliary variables that share a high degree of correlation with sensitive traits, researchers aim to bolster the reliability of RRT-based estimators and mitigate potential biases inherent in survey data.

These investigations delve into the theoretical foundations and practical implications of incorporating auxiliary information into quantitative RRT models, shedding

light on the efficacy and limitations of this approach. By examining the interaction between auxiliary variables and sensitive traits, researchers endeavor to identify optimal strategies for integrating auxiliary information into RRT methodologies, thereby maximizing the precision and validity of survey estimates.

The body of research surrounding quantitative RRT models in the presence of auxiliary information underscores the importance of methodological innovation and empirical inquiry in survey sampling. Through systematic analysis and empirical validation, scholars strive to refine existing methodologies and develop novel approaches that enhance the accuracy and reliability of survey estimates. By elucidating the role of auxiliary information in quantitative RRT models, these studies contribute to the advancement of survey methodology and the attainment of more robust and meaningful insights from survey data. Some of these studies are discussed below.

Ratio Estimator

Sousa et al. (2010) introduced a ratio estimator within the framework of the additive RRT model. Their approach involved estimating the mean of the sensitive variable through an enhanced estimator, utilizing a non-sensitive auxiliary variable under a non-optional RRT model. In their model, Y represents the sensitive study variable, which cannot be directly observed, while X stands for the non-sensitive auxiliary variable, positively correlated with Y. Additionally, S denotes a scrambling variable independent of both X and Y, with a mean of $\mu_S = 0$ and variance of σ_S^2. Let μ_X be the known true population mean and σ_X^2 be the known variance of the non-sensitive auxiliary variable X. Let μ_Y be the unknown true population mean and σ_Y^2 be the unknown variance of the sensitive study variable Y. For an additive RRT model, the respondent is asked to provide a scrambled response for Y given by $Z = Y + S$.

Assuming $E(S) = 0$, we get $E(Z) = E(Y)$ and the unbiased ordinary estimator under this RRT model is given by,

$$\mu_0 = \bar{z}. \tag{2.32}$$

The mean squared error (MSE) of the above basic estimator, under simple random sampling without replacement (SRSWOR), is given by

$$MSE(\hat{\mu}_0) = \frac{1-f}{n}(\sigma_Y^2 + \sigma_S^2), \tag{2.33}$$

where $f = \frac{n}{N}$ and N and n are the size of the finite population and the simple random sample drawn from it, respectively.

Based on this basic mean estimator, Sousa et al. (2010) proposed the ratio estimator for the mean of the sensitive variable Y, which is given by

$$\hat{\mu}_R = \bar{z}\big(\frac{\mu_X}{\bar{x}}\big), \tag{2.34}$$

where $\bar{z}$ is the sample mean of reported responses and $\bar{x}$ is the sample mean of the auxiliary variable.

The MSE of the above ratio estimator, under SRSWOR, is given by

$$MSE(\hat{\mu}_R) \approx \frac{1-f}{n}\mu_z^2(C_x^2 + C_z^2 - 2\rho_{zx}C_zC_x), \tag{2.35}$$

where $f = \frac{n}{N}$, $C_z = \frac{S_z}{\mu_z}$ and $C_x = \frac{S_x}{\mu_x}$ are the coefficients of variation of Z and X, respectively, and $\rho_{zx} = \frac{S_{zx}}{S_zS_x}$ is the coefficient of correlation between Z and X.

Sousa et al. (2010) also compared the efficiency of the ratio estimator, in terms

of its MSE in (2.35), with the MSE of the basic mean estimator from (2.33), it was established that the estimator proposed by Sousa et al. (2010) is more efficient.

It can be observed that $\text{MSE}(\hat{\mu}_R) < \text{MSE}(\hat{\mu}_0)$ if

$$\rho > \frac{1}{2}\frac{C_x}{C_y}\sqrt{1+\frac{\sigma_S^2}{\sigma_Y^2}}. \tag{2.36}$$

Regression Estimator

Gupta et al. (2012) proposed a regression estimator based on a non-optional RRT model and later extended it to a more general version. The basic regression estimator is given by

$$\hat{\mu}_{Reg} = \bar{z} + \hat{\beta_{zx}}(\mu_X - \bar{x}), \tag{2.37}$$

where $\hat{\beta_{zx}}$ is the sample regression coefficient between Z and X, and $Z = Y + S$ is the scrambled response on Y.

The MSE of the above estimator, under SRSWOR, is given by

$$MSE(\hat{\mu}_{Reg}) \approx \frac{1-f}{n}S_y^2\left[\left(1+\frac{S_s^2}{S_y^2}\right)-\rho_{yx}^2\right]. \tag{2.38}$$

It can be verified that:

- $\text{MSE}(\hat{\mu}_{Reg}) < \text{MSE}(\hat{\mu}_0)$ if

$$\rho_{yx}^2 > 0; \tag{2.39}$$

- $\text{MSE}(\hat{\mu}_{Reg}) < \text{MSE}(\hat{\mu}_R)$ if

$$(C_x - C_z\rho_{zx})^2 > 0. \tag{2.40}$$

These conditions will always hold true indicating that up to first order of approximation, the regression estimator $\hat{\mu}_{Reg}$ outperforms both the ordinary mean estimator $\hat{\mu}_0$ and the ratio estimator $\hat{\mu}_R$.

Gupta et al. (2014) further improved the above estimators (i.e. ratio estimator and basic regression estimator) by using optional RRT methodology.

Regression-Cum-Ratio Estimator

As mentioned previously, Gupta et al. (2012) later proposed a regression-cum-ratio estimator that combines both the regression estimator and ratio estimator. This estimator is given by

$$\hat{\mu}_{RCR} = [k_1\bar{z} + k_2(\mu_X - \bar{x})]\left(\frac{\mu_X}{\bar{x}}\right), \tag{2.41}$$

where k_1 and k_2 are constants.

The optimum values of k_1 and k_2 are given by

$$k_{1,opt} = \frac{1 - \frac{1-f}{n}C_x^2}{1 - \frac{1-f}{n}\big(C_x^2 - C_z^2(1-\rho_{zx}^2)\big)} \tag{2.42}$$

and

$$k_{2,opt} = \frac{\mu_Y}{\mu_X}\left(1 + k_{1,opt}\big(\frac{\rho_{zx}C_z}{C_x} - 2\big)\right). \tag{2.43}$$

Substituting the optimum values of k_1 and k_2 in (2.42) and (2.43), we get

$$MSE(\hat{\mu}_{RCR})_{min} \approx \frac{\frac{1-f}{n}\bar{Y}^2C_z^2(1-\rho_{zx}^2)(1-\frac{1-f}{n}C_x^2)}{\frac{1-f}{n}C_z^2(1-\rho_{zx}^2) + (1-\frac{1-f}{n}C_x^2)} \tag{2.44}$$

It can be verified that

- $\text{MSE}(\hat{\mu}_{RCR})_{min} < \text{MSE}(\hat{\mu}_0)$ if

$$\frac{1-f}{n}(S_y^2 + S_s^2) > 0; \tag{2.45}$$

- $\text{MSE}(\hat{\mu}_{RCR})_{min} < \text{MSE}(\hat{\mu}_R)$ if

$$\left(\frac{C_x}{C_z} - \rho_{zx}\right)^2 + \frac{\frac{1-f}{n}C_z^2(1-\rho_{zx}^2)^2}{\frac{1-f}{n}C_z^2(1-\rho_{zx}^2) + (1-\frac{1-f}{n}C_x^2)} > 0. \tag{2.46}$$

- $\text{MSE}(\hat{\mu}_{RCR})_{min} < \text{MSE}(\hat{\mu}_{Reg})$ if

$$\frac{1-f}{n}C_z^2(1-\rho_{zx}^2) > 0. \tag{2.47}$$

Based on these conditions, it can be inferred that the generalized regression-cum-ratio estimator $\hat{\mu}_{RCR}$ outperforms not only the basic mean estimator $\hat{\mu}0$ and the regression estimator $\hat{\mu}Reg$ but also the ratio estimator $\hat{\mu}_R$ when $\frac{1-f}{n}C_x^2 < 1$.

Generalized Estimator

Khalil, Noor-ul Amin & Hanif (2018) introduced a generalized mean estimator for sensitive variables with a robust approach that can effectively account for measurement errors on both the reported response Z and the non-sensitive auxiliary variable X which is positively correlated with the sensitive study variable Y. This generalized estimator is given by

$$\hat{\mu}_{Gen} = [\bar{z} + k(\bar{X} - \bar{x})]\left[\frac{a\bar{X} + b}{\lambda(a\bar{x} + b) + (1-\lambda)(a\bar{X} + b)}\right]^v, \tag{2.48}$$

where k and v are suitable constants and λ is an unknown constant determined through optimality considerations.

In (2.48), assuming $a(\neq 0)$ and b as known parameters of the auxiliary variable X, diverse estimators can be derived by employing different values of these parameters. These parameters include the coefficient of variation (C_x), population correlation coefficient (ρ_{zx}), coefficient of skewness $(\beta_1(x))$, etc. Moreover, setting $v = 1$ leads to various ratio estimators, while $v = -1$ results in product estimators. This inherent flexibility of the generalized mean estimator enriches the applicability of the estimation approach.

Let the measurement errors associated with the reported response Z and the non-sensitive auxiliary variable X respectively be given by $U_i = z_i - Z_i$ and $V_i = x_i - X_i$. These measurement errors are assumed to be random and uncorrelated with mean zero and variances S_U^2 and S_V^2, respectively.

With the optimum value of λ,

$$MSE(\hat{\mu}_{Gen})_{min} \approx \theta\left(S_Z^2 + S_U^2 - \frac{\rho_{ZX}^2 S_Z^2 S_X^2}{S_X^2 + S_V^2}\right). \tag{2.49}$$

It can be verified that when measurement errors are present,

- $\text{MSE}(\hat{\mu}_{Gen})_{min} < \text{MSE}^*(\hat{\mu}_0)$ if

$$\rho_{ZX}^2 \frac{S_Z^2 S_X^2}{S_X^2 + S_V^2} > 0; \tag{2.50}$$

- $\text{MSE}(\hat{\mu}_{Gen})_{min} < \text{MSE}^*(\hat{\mu}_R)$ if

$$\left(\frac{\alpha \bar{Z}}{\alpha \bar{X} + \beta}\sqrt{S_X^2 + S_V^2} - \frac{\rho_{ZX} S_Z S_X}{\sqrt{S_X^2 + S_V^2}}\right)^2 > 0. \tag{2.51}$$

2.2 Estimating the Distribution of a Study Variable

As discussed in Section 1.2, density estimation emerges as a valuable tool in the analysis of sensitive variables. By capturing the intricate nuances of the variable of interest, density estimation furnishes researchers with a comprehensive understanding of its underlying distribution through the depiction of its potential density curve. This broader information encompasses not only the central tendency but also the variability and shape of the distribution, offering insights that extend beyond mere point estimates.

Furthermore, the utility of density estimation extends beyond descriptive purposes, facilitating the subsequent inference of population parameters associated with the variable under study. Through the estimated density curve, researchers can derive a plethora of population characteristics, including moments, quantiles, and other key descriptors. This wealth of information enables a deeper exploration of the variable's properties and empowers researchers to make informed decisions and draw meaningful conclusions from the data.

In essence, density estimation serves as a cornerstone in the analytical toolkit of researchers, providing a multifaceted lens through which to examine sensitive variables. By illuminating the distributional characteristics of the variable of interest, density estimation fosters a deeper understanding of its behavior and informs subsequent statistical analyses and inferential procedures.

2.2.1 Density Estimation in the Presence of Auxiliary Information

The integration of auxiliary information has emerged as a promising avenue for enhancing the accuracy and precision of density estimators. While existing research has yet to specifically address the estimation of density curves for sensitive variables in conjunction with auxiliary information, studies have explored the utilization of auxiliary data to bolster density estimation for non-sensitive variables. These include Rao et al. (1990), Dubnicka (2009), and Mostafa & Ahmad (2019).

In these investigations, researchers have examined the efficacy of incorporating auxiliary information into the density estimation process for general non-sensitive variables. By leveraging auxiliary data sources that provide additional insights into the underlying distribution, researchers have sought to refine and optimize density estimators, thereby improving their performance and reliability.

The rationale behind incorporating auxiliary information lies in its potential to complement and enrich the information gleaned from the primary dataset. Auxiliary data, which may encompass demographic characteristics, contextual variables, or external covariates, offer valuable insights into the underlying structure of the data, enhancing the precision and robustness of density estimators.

While the specific application of auxiliary information to the estimation of density curves for sensitive variables remains unexplored in the existing literature, the principles and methodologies established in studies on non-sensitive variables offer valuable insights and potential avenues for future research. By adapting and extending existing techniques for incorporating auxiliary information, researchers may develop innovative approaches tailored to the unique challenges and considerations inherent in estimating

density curves for sensitive variables.

Estimation of Population Distribution Functions and Quantiles Using Auxiliary Information

Rao et al. (1990) explores the efficiency and accuracy of various estimators for population distribution functions and quantiles under different sampling designs, focusing on both design-based and model-based approaches. Initially, it highlights the limitations of customary design-based estimators compared to model-based ones, particularly in utilizing auxiliary population information. The study then introduces design-based ratio and difference estimators, contrasting them with a model-based method proposed by Chambers & Dunstan (1986). Through simulation studies using populations of sugar cane farms and synthetic datasets, the paper demonstrates the advantages of design-based estimators over model-based ones, especially under model mis-specifications and for large samples.

The results reveal that design-based estimators, particularly the difference estimators, exhibit superior performance in terms of relative mean errors and root mean square errors compared to model-based estimators under various scenarios. The study emphasizes the importance of incorporating auxiliary information in estimating population distribution functions and quantiles, showing that design-based estimators offer greater robustness and efficiency, especially for larger sample sizes. Additionally, the article discusses variance estimation techniques for the proposed estimators and highlights the potential for further extensions to handle more complex sampling designs and multiple auxiliary variables. Overall, the findings underscore the significance of adopting appropriate estimation methods that leverage auxiliary information effectively to enhance the accuracy and efficiency of population distribution function and

quantile estimates.

Kernel Density Estimation With Missing Data Using Auxiliary Information

Dubnicka (2009) proposed a method for estimating the density of a response variable, possibly missing observations at random, with available auxiliary data. The kernel density estimator is based on the Horvitz-Thompson estimator, assuming the missingness of the response variable at random. Simulation studies demonstrated the superior performance of the proposed density estimator in terms of integrated squared error (ISE) compared to the complete-case density estimator. Moreover, it performed nearly as well as the density estimator utilizing the full dataset, as if no values were missing.

The authors preferred the modified Horvitz-Thompson-type kernel density estimator initially by Horvitz & Thompson (1952) with Nadaraya-Watson estimates by (Nadaraya 1964) and Watson (1964) of propensity scores over other Horvitz-Thompson-type density estimators due to its nature as a density function and the relative ease of obtaining Nadaraya-Watson probability estimates. Additionally, the Sheather-Jones bandwidth selection procedure based on observed responses exhibited the best performance across various true densities.

In practical applications, the modified Horvitz-Thompson-type kernel density estimator effectively adjusted the complete-case kernel density estimator to accommodate scenarios where HIV patients with low CD4 counts are more likely to miss visits later in the study due to declining health. This suggests that when auxiliary variables highly correlated with the response are available, this modified kernel density estimator may be suitable even when the response variable is not missing at random.

Areas for further investigation include bandwidth selection for the Nadaraya-Watson estimator of propensity scores and developing methods to test for the equality

of densities. Future research aims to extend methods for estimating the conditional density of the response variable given auxiliary variables when responses are not completely observed. This includes testing for the equality of densities to determine differences among treatment groups and extending improved conditional density estimators to scenarios where the response variable is missing at random, along with addressing issues of bandwidth selection.

Model-Assisted Kernel Density Estimation in the Presence of Auxiliary Information

Suppose the non-sensitive study variable Y is observed in s units and is predictable in $\bar{s}$ units in a finite population U. Suppose the functional relationship between Y and an auxiliary variable X can be described by a parametric regression model as follows:

$$y_i = \mu(x_i, \boldsymbol{\beta}) + \sigma(x_i)\epsilon_i; i \in U, \tag{2.52}$$

where

- $\mu(\cdot, \cdot)$ - known mean function;
- $\sigma(\cdot, \cdot)$ - known strictly positive function;
- $\boldsymbol{\beta}$ - unknown model parameter vector;
- ϵ - i.i.d. with zero mean and unit variance.

Mostafa & Ahmad (2019) proposed the following model-assisted kernel density

estimator for $f(y)$:

$$\hat{f}_{par}(y;h) = \frac{1}{N}\big[\sum_{i\in s} d_i\{K_h(y-y_i) - K_h(y-\hat{y}_i)\} + \sum_{i\in U} K_h(y-\hat{y}_i)\big], \tag{2.53}$$

where $\hat{y}_i = \mu(x_i, \hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\beta}}$ is the design-weighted least squares estimate of $\boldsymbol{\beta}$ using the sampling weights d_i. $K(x)$ is a known kernel, typically a symmetric and bounded probability density function (pdf). The bandwidth, denoted as h, is a positive value that determines the smoothness of the density curve.

The estimator proposed in the study demonstrated commendable performance compared to traditional kernel density estimators (KDEs) that do not take into account auxiliary information. This superiority was particularly notable under circumstances where the relationship between the auxiliary variable and the primary study variable was accurately modeled.

The empirical findings underscored the efficacy of the proposed estimator in leveraging auxiliary information to enhance the accuracy and precision of density estimation. By incorporating auxiliary variables that capture additional dimensions of the data, the proposed estimator was able to capture more nuanced patterns and trends in the underlying distribution, resulting in superior performance compared to traditional KDEs.

2.2.2 Density Estimation Under a Multiplicative RRT Model

Poole (1974) pioneered the idea of estimating the distribution function of a continuous type random variable using a variation of Warner's linear randomized response multiplicative RRT models in Warner (1971). Poole's technique, illustrated through the estimation of income distribution from a sample of 500 individuals, showcased

the potential use of such methods in maintaining data confidentiality. Warner's randomized response procedure, where respondents answer one of two randomly chosen questions to ensure anonymity, laid the groundwork for Poole's approach, which involved concealing true responses by multiplying them with random numbers.

The methodology presented in Poole's work extends the randomized response technique to estimate the entire distribution of a quantitative variable, not just its mean and variance. Through theoretical development, Poole (1974) established a method to estimate distribution functions of continuous variables by concealing respondents' true answers and using known properties of the distribution of random multipliers.

The estimation procedure outlined by Poole (1974) involves survey participants multiplying their responses by random numbers, thereby concealing their true answers. Utilizing the recorded randomized responses, along with known properties of the random multiplier distribution, Poole's method enables the estimation of distribution functions for sensitive quantitative variables. The article provides both theoretical underpinnings and practical implementation steps, offering a valuable contribution to the literature on survey methodology and data privacy.

2.2.3 Analysis and Optimization of Kernel Density Estimation

Wand & Jones (1994) discussed the principles, applications, and analysis of kernel smoothers. For a general kernel density estimate $\hat{f}_h(x)$ in (1.1), its mean squared error (MSE) at a single point x_0 is given by

$$MSE(\hat{f}_h(x_0)) = \frac{h^4 c_K^2 \big(f''(x_0)\big)^2}{4} + \frac{f(x_0)}{nh} \int K^2(y)dy + O(h^4) + O(\frac{1}{nh}), \qquad (2.54)$$

where $K(y)$ is a pre-determined kernel function, h is the bandwidth, and $c_K = \int y^2 K(y)dy$.

Then the mean integrated square (MISE), i.e. the overall MSE of entire function, is given by

$$\text{MISE}(\hat{f}_h(x)) = \text{AMISE}(\hat{f}_h(x)) + O(h^4) + O(\frac{1}{nh}), \tag{2.55}$$

where AMISE (asymptotic MISE) $= \frac{h^4 c_K^2 \int (f''(x_0))^2 dx}{4} + \frac{\int K^2(y)dy}{nh}$.

The optimal bandwidth h_{opt} that minimizes the above AMISE is given by

$$h_{opt} = \left(\frac{\int K^2(y)dy}{nc_K^2 \int (f''(x_0))^2 dx} \right)^{1/5} \tag{2.56}$$

2.2.4 Kernel Density Estimation Under a Multiplicative RRT Model

Referring to Poole (1974) and Wand & Jones (1994), Ahmad (2002) later introduced the kernel estimation of the density curve of a sensitive variable based on a multiplicative RRT model with partial theoretical results.

Let

$Y =$ *Sensitive study variable*

$S =$ *Scrambling variable (independent of* Y*;* $S \sim Uniform(0, T), T > 0$

$n =$ *Sample size*

$h =$ *Bandwidth*

$Z =$ *Reported response, where* $Z = Y \cdot S$

$F(s), G(y), Q(z) =$ *Cumulative distribution function (CDF) of S, Y, and Z, respectively*

$f(s), g(y), q(z)$ = *Probability density function (PDF) of S, Y, and Z, respectively*

Ahmad (2002) proposed the following kernel density estimator using multiplicative RRT models:

$$\hat{g}_M(y) = -yT^2\hat{q}'_{KDE}(yT), \tag{2.57}$$

where $\hat{q}_{KDE}(z) = (1/nh)\sum_{i=1}^{n} K[(z - Z_i)/h]$.

With the first-order approximation, the approximate MSE (AMSE) of $\hat{g}_M(y)$ is given by

$$\text{AMSE}(\hat{g}_M(y)) = \frac{y^2T^4q(yT)}{nh^3}R(K') + \frac{c_K^2h^4}{4}y^2T^4\left(q'''(yT)\right)^2, \tag{2.58}$$

where $R(K') = \int K'^2(x)dx$, and $c_K = \int x^2K(x)dx$.

In the special case where scrambling variable S follows $Uniform(0, 1)$, its approximate mean integrated squared error (AMISE) is given by

$$\text{AMISE}(\hat{g}_M(y)) = \frac{R(K')\int y^2q(y)dy}{nh^3} + \frac{c_K^2h^4R(\phi_q''')}{4}, \tag{2.59}$$

where $\phi_q''' = yq'''(y)$.

The minimized AMISE$(\hat{g}_M(y))$ occurs at

$$h_{opt,M} = \left(\frac{R(K')\int y^2q(y)dy}{c_KR(\phi_q''')n}\right)^{1/7}. \tag{2.60}$$

This $h_{opt,M}$ value can be considered as the optimal bandwidth.

Chapter 3: Kernel Density Estimation Using Additive Randomized Response Technique (RRT) Models

3.1 Introduction

Existing research on RRT models, such as Gupta et al. (2012), indicates a prevalent preference for additive scrambling over multiplicative scrambling, owing to its various advantages. Notably, additive RRT models outperform multiplicative RRT models in terms of mean estimators, while also providing better privacy protection. For example, in multiplicative RRT models, a non-zero scrambled response inherently reveals that the true response cannot be zero, thereby compromising privacy. Furthermore, the intuitive nature of additive models renders them more accessible, particularly for survey respondents with limited mathematical proficiency.

Motivated by these insights, this Chapter centers on the transition from employing kernel density estimation (KDE) for multiplicative RRT to the additive RRT model. In this Chapter, we build upon the pioneering work of Ahmad (2002) by extending

and validating the theoretical framework of KDE within a more general context. We investigate the application of KDE using an additive RRT model and propose a corresponding KDE estimator. We then proceed to derive expressions for both the mean square error (MSE) and its integrated form (MISE) for this proposed estimator, ensuring accuracy up to the first order of approximation. Furthermore, we conduct a comprehensive performance evaluation by comparing the efficacy of our proposed estimator with that of the kernel density estimator relying on multiplicative RRT models.

3.2 Review Kernel Density Estimation Under a Multiplicative RRT Model

As discussed in Section 2.2, Ahmad (2002) introduced the kernel estimation of the density curve of a sensitive variable based on multiplicative RRT models with some theoretical results. Ahmad's pioneering work laid the theoretical groundwork for this innovative methodology, providing initial insights into its feasibility and efficacy in capturing the underlying distribution of sensitive variables.

Building upon Ahmad's seminal contributions, our research endeavors to extend and generalize the theoretical framework established in his work to accommodate a broader spectrum of scenarios. By relaxing certain constraints and allowing for increased variability or noise within the model, we aim to enhance its applicability and robustness across diverse survey settings.

To validate the extended theoretical framework, we conducted a comprehensive simulation study, wherein the proposed methodology was subjected to rigorous testing

under various conditions and scenarios. Through meticulous analysis and empirical validation, we sought to ascertain the validity and reliability of the theoretical results, thus bolstering confidence in the efficacy of the proposed approach.

3.2.1 Extending and Validating Multiplicative Kernel Density Estimator

For multiplicative scrambling models, the respondent is asked to provide a scrambled response for Y given by

$$Z_M = Y \cdot S, \tag{3.1}$$

where Y is the sensitive study variable, and the scrambling variable $S \sim U(0, T)$ is independent of Y. Let $F(s), G(y), Q(z)$ denote the CDF of S, Y, Z, respectively, with corresponding pdf $f(s), g(y)$, and $q(z)$.

During the simulation process, it was detected that slight modifications may be required in 2.60 to calculate the optimal bandwidth h^*. The modified version is shown as follows:

$$h^*_{opt,M} = \Big(\frac{3R(K')\int y^2 q(y)dy}{c_K^2 R(\phi_q''')n}\Big)^{1/7}, \tag{3.2}$$

where $R(K') = \int K'^2(x)dx$, $c_K = \int x^2 K(x)dx$, and $\phi_q''' = yq'''(y)$.

From Table 3.1, the theoretical asymptotic MISE (AMISE) of the special case $T = 1$ by Ahmad (2002) is a reasonably good match to the empirical MISE values. This correspondence between theoretical results and empirical observations suggests that the theoretical framework proposed by Ahmad (2002) effectively captures the

underlying dynamics of the multiplicative RRT model.

Table 3.1. Theoretical AMISEs and empirical MISEs of multiplicative kernel density estimator with $h = 0.4, 0.6, 0.8$, and optimal value for $T = 1$.

		h	MISE_E	MISE_T
	100	0.4	0.2363	0.2247
n	500	0.4	0.0497	0.0452
	1000	0.4	0.0273	0.0228
	100	0.6	0.0724	0.0701
n	500	0.6	0.0186	0.0176
	1000	0.6	0.0136	0.0110
	100	0.8	0.0382	0.0332
n	500	0.8	0.0143	0.0108
	1000	0.8	0.0116	0.0080
		$h^*_{opt,M}$	MISE_E	MISE_T
	100	0.9770	0.0265	0.0269
n	500	0.7766	0.0146	0.0167
	1000	0.7034	0.0120	0.0072

Moreover, our analysis reveals that employing the optimal bandwidth, as defined in (3.2), consistently yields lower MISE values across most cases. This finding underscores the importance of selecting an appropriate bandwidth parameter in kernel density estimation, as it directly influences the accuracy and precision of the density estimate. By optimizing the bandwidth parameter according to the proposed formula, researchers can achieve superior performance in estimating the density curve of sensitive variables within the context of multiplicative RRT models.

We also extended 2.59 of the special case $T = 1$ to a more general case where T is any positive number. The generalized AMISE is given by

$$\text{AMISE}(\hat{g}_M(y)) = \frac{T^2 R(K') \int y^2 q(y) dy}{nh^3} + \frac{T^2 c_K^2 h^4 R(\phi^{'''}_{q,T})}{4}, \tag{3.3}$$

where $\phi^{'''}_{q,T} = yTq^{'''}(yT)$.

3.3 Proposed Kernel Density Estimator Under an Additive RRT Model

We propose a kernel density estimator in the context of additive RRT models. For additive scrambling models, the respondent is asked to provide a scrambled response for Y given by

$$Z_A = Y + S, \tag{3.4}$$

where Y is the sensitive study variable, and the scrambling variable $S \sim U(0, T)$ is independent of Y. Let $F(s), G(y), Q(z)$ denote the CDF of S, Y, Z, respectively, with corresponding pdf $f(s), g(y)$, and $q(z)$.

Since $S \sim U(0, T)$, $F(s) = s/T, s \in [0, T]$.

$$\begin{aligned} G(y) &= P(Z - S \leq y) \\ &= \int_0^{\infty} q(z)[1 - F(z - y)]dz \\ &= \int_y^{y+T} dQ(z) - \frac{1}{T}\int_y^{y+T}(z - y)q(z)dz + Q(y) - Q(0) \\ &= Q(y + T) - Q(0) - \frac{1}{T}\int_y^{y+T}(z - y)q(z)dz. \end{aligned} \tag{3.5}$$

Using KDE, $G(y)$ can be estimated by

$$\hat{G}(y) = \hat{Q}_{KDE}(y + T) - \hat{Q}_{KDE}(0) - \frac{1}{T}\int_y^{y+T}(z - y)\hat{q}_{KDE}(z)dz, \tag{3.6}$$

where $\hat{Q}_{KDE}(z) = \int_{-\infty}^{z} \hat{q}_{KDE}(w)dw$, and $\hat{q}_{KDE}(z) = (1/nh)\sum_{i=1}^{n} K[(z - Z_i)/h]$ with

the kernel $K(x)$ and the bandwidth h.

Taking the derivative of both sides of (3.6), the proposed kernel density estimator is given by

$$\begin{aligned}\hat{g}_A(y) &= \frac{1}{T}\Big(\hat{Q}_{KDE}(y+T) - \hat{Q}_{KDE}(y)\Big) \\ &= \frac{1}{T}\int_y^{y+T} \hat{q}_{KDE}(w)dw. \end{aligned} \tag{3.7}$$

3.4 Efficiency and Bandwidth Selection for Kernel Density Estimator with Additive RRT Models

To estimate the mean integrated square error (MISE) of the proposed kernel density estimator $\hat{g}_A(y)$, we first need to find the mean square error (MSE) of $\hat{g}_A(y)$ at any point y as follows:

$$\begin{aligned}\mathrm{MSE}(\hat{g}_A(y)) &= E\Big(\hat{g}_A(y) - g(y)\Big)^2 \\ &= E\Big(\frac{1}{T}\int_y^{y+T} \hat{q}(w)dw - g(y)\Big)^2 \\ &= \frac{1}{T^2}\int_y^{y+T} \mathrm{MSE}(\hat{q}(w))dw + E\Big(\frac{1}{T}\int_y^{y+T} q(w)dw - g(y)\Big)^2 \\ &= \frac{1}{T^2}\int_y^{y+T} \mathrm{MSE}(\hat{q}(w))dw + \frac{1}{T^2}\big(\int_y^{y+T} q(w)dw\big)^2 + \big(g(y)\big)^2 \\ &\quad - \frac{2g(y)}{T}\int_y^{y+T} q(w)dw. \end{aligned} \tag{3.8}$$

Wand & Jones (1994) derived the approximate MSE (AMSE) expression for kernel density estimation in a general sense. Thus, for Model (3.4) the AMSE($\hat{q}(w)$) can be

expressed as follows:

$$\text{AMSE}(\hat{q}(w)) = \frac{h^4 c_K^2 \big(q''(w)\big)^2}{4} + \frac{q(w)}{nh} \int \big(K(y)\big)^2 dy, \tag{3.9}$$

where $c_K = \int x^2 K(y) dy$.

From (3.8) and (3.9), using the Taylor's approximation and retaining terms of order up to 2, the AMSE and the approximate mean integrated squared error (AMISE) of $\hat{g}_A(y)$ are given by

$$\begin{aligned}\text{AMSE}(\hat{g}_A(y)) = {} & \frac{h^4 c_K^2}{4T^2} \int_y^{y+T} \big(q''(w)\big)^2 dw + \frac{\int \big(K(y)\big)^2 dy}{nhT^2} \int_y^{y+T} q(w) dw \\ & + \frac{1}{T^2} \Big(\int_y^{y+T} q(w) dw \Big)^2 + \big(g(y)\big)^2 - \frac{2g(y)}{T} \int_y^{y+T} q(w) dw \end{aligned} \tag{3.10}$$

and

$$\begin{aligned}\text{AMISE}(\hat{g}_A(y)) = {} & \frac{h^4 c_K^2}{4T^2} \iint_y^{y+T} \big(q''(w)\big)^2 dw dy + \frac{\int \big(K(y)\big)^2 dy}{nhT^2} \iint_y^{y+T} q(w) dw dy \\ & + \frac{1}{T^2} \int \Big(\int_y^{y+T} q(w) dw \Big)^2 dy + \int \big(g(y)\big)^2 dy \\ & - \frac{2}{T} \int g(y) \int_y^{y+T} q(w) dw dy. \end{aligned} \tag{3.11}$$

Differentiating (3.11) with respect to h we get the following optimal value of h:

$$h_{opt,A} = \left(\frac{\int K^2(y) dy}{n c_K^2} \cdot \frac{\iint_y^{y+T} q(w) dw dy}{\iint_y^{y+T} \big(q''(w)\big)^2 dw dy} \right)^{1/5}, \tag{3.12}$$

which minimizes the AMISE.

3.5 Simulation Study

In this section, we present results of a simulation study with particular focus on the performance of the proposed kernel density estimator $\hat{g}_A(y)$ using additive scrambling as compared to the kernel density estimator $\hat{g}_M(y)$ using multiplicative scrambling.

In the simulation study, we consider a finite population of size $N = 5,000$ generated from a normal distribution with the mean $\mu_Y = 5$ and variance $\sigma_Y^2 = 5$. The scrambling variable S is taken to be a uniform variate from $U(0, T)$, where $T = 1, 5, 10$. For a multiplicative RRT model, the reported response is given by $Z_M = Y \cdot S$, and for an additive RRT model, the reported response is given by $Z_A = Y + S$.

For both models mentioned above, we choose the normal kernel when using KDE, which means $k(x) = \chi(x)$, and χ is the standard normal density function. The simulation study can be divided into two parts. In the first part, the bandwidth h is pre-defined as $h = 0.4, 0.6, 0.8$ during the KDE process so that we can observe the relationship between MISE and the sample size n, the scrambling scale T and the bandwidth h. In the second part, we use the optimal bandwidth h_{opt} obtained from (3.2) and (3.12) that would theoretically minimize AMISE so as to compare the performance of the density estimators.

3.5.1 Comparison of Kernel Density Estimators

We consider three sample sizes: $n = 100, 500, 1000$, using SRSWOR (simple random sampling without replacement). Coding for the simulations was done in R and results are averaged over 1,000 iterations. The empirical MISE of the kernel density estimator

$\hat{g}(y)$ is computed by

$$\text{MISE}_E(\hat{g}(y)) = \frac{1}{1000}\sum_{i=1}^{1000}\int \big(\hat{g}(y) - g(y)\big)^2 dy,$$

where $\hat{g}(y) = \{\hat{g}_M(y), \hat{g}_A(y)\}$.

Tables 3.2 and 3.3 provide valuable insights into the comparison between theoretical AMISEs and empirical MISEs for kernel density estimators employing various RRT models. A notable observation is the close match between the theoretical AMISEs and the empirical MISEs, indicating a robust alignment between theoretical expectations and real-world performance across different RRT models.

Upon closer examination, it becomes apparent that instances characterized by $T = 0$ typically exhibit the lowest MISE values, indicative of superior estimation accuracy under minimal noise conditions. As the value of T increases, the MISE values for the kernel density estimators tend to ascend correspondingly. This trend holds intuitive appeal, as higher values of T correspond to greater levels of noise introduced into respondents' true responses. Consequently, the increasing magnitude of MISEs with higher T values reflects the deleterious impact of heightened noise levels on the accuracy and precision of estimation.

This observed trend underscores the importance of judiciously selecting the level of noise introduced in RRT models, striking a delicate balance between privacy preservation and estimation accuracy. While the $T = 0$ case offers the advantage of no noise and consequently lower MISE values, higher T values necessitate a trade-off between increased privacy protection and compromised estimation quality.

In Table 3.2, we present the outcomes obtained from the multiplicative kernel density estimator $\hat{g}_M(y)$ alongside the proposed additive kernel density estimator

$\hat{g}_A(y)$, employing bandwidth values of $h = 0.4, 0.6, 0.8$. The results unveil a consistent trend wherein the proposed additive kernel density estimator consistently outperforms its multiplicative counterpart, as evidenced by its consistently smaller MISE values across varying bandwidth settings.

It is worth noting that slight adjustments to the bandwidth parameter can exert substantial influence on the MISE values of the multiplicative kernel density estimator $\hat{g}_M(y)$, while exerting comparatively less impact on the MISE values of the additive kernel density estimator $\hat{g}_A(y)$. For example, in scenarios characterized by $T = 1$ and $n = 500$, the MISE of $\hat{g}_A(y)$ exhibits relative stability even as the bandwidth decreases from 0.8 to 0.6 and further to 0.4. In stark contrast, the MISE of $\hat{g}_M(y)$ demonstrates significant fluctuations, escalating approximately 15.9 times, 23.3 times, and 49.7 times higher than those of $\hat{g}_A(y)$ correspondingly. This observed trend persists across all analyzed cases, underscoring the robustness of the proposed additive kernel density estimator to variations in bandwidth selection.

Furthermore, the presence of NAs in Table 3.2 underscores the potential pitfalls associated with suboptimal bandwidth selection in multiplicative RRT-based KDEs, as it may impede the effectiveness of density estimation.

Table 3.2. Theoretical (**bold**) AMISEs and empirical MISEs of the kernel density estimators with $h = 0.4, 0.6, 0.8$. $T = 1$.

n	h	Original KDE Without Scrambling (T=0)	T=1 Var(S)=0.08 $\hat{g}_A(y)$	T=1 Var(S)=0.08 $\hat{g}_M(y)$
100	0.4	0.0059	0.0046	0.2363
		0.0071	**0.0071**	**0.2247**
	0.6	0.0037	0.0033	0.0724
		0.0048	**0.0048**	**0.0701**
	0.8	0.0027	0.0027	0.0382
		0.0039	**0.0039**	**0.0332**
500	0.4	0.0012	0.0010	0.0497
		0.0014	**0.0015**	**0.0452**
	0.6	0.0008	0.0008	0.0186
		0.0010	**0.0011**	**0.0176**
	0.8	0.0007	0.0009	0.0143
		0.0010	**0.0011**	**0.0108**
1000	0.4	0.0006	0.0005	0.0273
		0.0007	**0.0007**	**0.0228**
	0.6	0.0004	0.0005	0.0136
		0.0006	**0.0006**	**0.0110**
	0.8	0.0004	0.0006	0.0116
		0.0007	**0.0007**	**0.0080**

Table 3.2 (Continued). Theoretical (**bold**) AMISEs and empirical MISEs of the kernel density estimators with $h = 0.4, 0.6, 0.8$. $T = 5$.

n	h	Original KDE Without Scrambling (T=0)	T=5 Var(S)=2.08	
			$\hat{g}_A(y)$	$\hat{g}_M(y)$
100	0.4	0.0059	0.0086	27.1744
		0.0071	**0.0088**	**27.5722**
	0.6	0.0037	0.0089	8.2352
		0.0048	**0.0083**	**8.3415**
	0.8	0.0027	0.0095	3.4824
		0.0039	**0.0081**	**3.5280**
500	0.4	0.0012	0.0077	5.4818
		0.0014	**0.0076**	**5.9936**
	0.6	0.0008	0.0081	1.6148
		0.0010	**0.0075**	**1.7135**
	0.8	0.0007	0.0087	0.6934
		0.0010	**0.0075**	**0.7366**
1000	0.4	0.0006	0.0076	2.6971
		0.0007	**0.0075**	**3.2410**
	0.6	0.0004	0.0080	0.8087
		0.0006	**0.0075**	**0.9696**
	0.8	0.0004	0.0086	0.3474
		0.0007	**0.0074**	**0.4050**

Table 3.2 (Continued). Theoretical (**bold**) AMISEs and empirical MISEs of the kernel density estimators with $h = 0.4, 0.6, 0.8$. $T = 10$.

n	h	Original KDE Without Scrambling (T=0)	T=10 Var(S)=8.33 $\hat{g}_A(y)$	T=10 Var(S)=8.33 $\hat{g}_M(y)$
100	0.4	0.0059	0.0341	NA
		0.0071	**0.0342**	**228.3521**
	0.6	0.0037	0.0344	65.6665
		0.0048	**0.0340**	**67.5006**
	0.8	0.0027	0.0347	27.6782
		0.0039	**0.0339**	**28.4126**
500	0.4	0.0012	0.0337	NA
		0.0014	**0.0336**	**48.8122**
	0.6	0.0008	0.0340	12.9601
		0.0010	**0.0336**	**14.3035**
	0.8	0.0007	0.0343	5.4674
		0.0010	**0.0336**	**5.9701**
1000	0.4	0.0006	0.0337	NA
		0.0007	**0.0336**	**26.3697**
	0.6	0.0004	0.0340	6.3851
		0.0006	**0.0336**	**7.6539**
	0.8	0.0004	0.0343	2.6975
		0.0007	**0.0335**	**3.1648**

Turning to Table 3.3, the findings confirm the superior efficiency of the proposed additive kernel density estimator $\hat{g}_A(y)$ over its multiplicative counterpart $\hat{g}_M(y)$. Notably, for $T = 10$, the MISE of $\hat{g}_M(y)$ exceeds that of $\hat{g}_A(y)$ by approximately 23.7 times, 9.7 times, and 6.8 times for sample sizes of $n = 100$, $n = 500$, and $n = 1000$ respectively. Regardless of the value of T, the additive RRT-based kernel density estimator consistently achieves a lower minimum MISE compared to its multiplicative counterpart, underscoring its enhanced performance across diverse sample sizes and scenarios.

Table 3.3. Theoretical (**bold**) AMISEs and empirical MISEs of the kernel density estimators using the optimal bandwidth. $T = 1$.

	Original KDE Without Scrambling (T=0)			T=1 Var(S)=0.08	
n	h_{opt}	MISE	Estimator	h_{opt}	MISE
100	0.9681	0.0024	$\hat{g}_A(y)$	0.9758	0.0025
		0.0036			**0.0036**
		0.0036	$\hat{g}_M(y)$	1.1110	0.0222
					0.0164
500	0.7016	0.0007	$\hat{g}_A(y)$	0.7072	0.0008
		0.0010			**0.0010**
		0.0010	$\hat{g}_M(y)$	0.8828	0.0137
					0.0085
1000	0.6108	0.0004	$\hat{g}_A(y)$	0.6157	0.0005
		0.0006			**0.0006**
		0.0006	$\hat{g}_M(y)$	0.7996	0.0115
					0.0066

Table 3.3 (Continued). Theoretical (**bold**) AMISEs and empirical MISEs of the kernel density estimators using the optimal bandwidth. $T = 5$.

	Original KDE Without Scrambling (T=0)			T=5 Var(S)=2.08	
n	h_{opt}	MISE	Estimator	h_{opt}	MISE
100	0.9681	0.0024	$\hat{g}_A(y)$	1.1634	0.0109
					0.0079
		0.0036	$\hat{g}_M(y)$	2.0306	0.2184
					0.2206
500	0.7016	0.0007	$\hat{g}_A(y)$	0.8432	0.0089
					0.0075
		0.0010	$\hat{g}_M(y)$	1.6135	0.0877
					0.0939
1000	0.6108	0.0004	$\hat{g}_A(y)$	0.7341	0.0084
					0.0074
		0.0006	$\hat{g}_M(y)$	1.4614	0.0604
					0.0701

Table 3.3 (Continued). Theoretical (**bold**) AMISEs and empirical MISEs of the kernel density estimators using the optimal bandwidth. $T = 10$.

	Original KDE Without Scrambling (T=0)			T=10 Var(S)=8.33	
n	h_{opt}	MISE	Estimator	h_{opt}	MISE
100	0.9681	0.0024	$\hat{g}_A(y)$	1.6232	0.0368
					0.0337
		0.0036	$\hat{g}_M(y)$	2.5378	0.8712
					0.9007
500	0.7016	0.0007	$\hat{g}_A(y)$	1.1764	0.0351
					0.0336
		0.0010	$\hat{g}_M(y)$	2.0165	0.3421
					0.3702
1000	0.6108	0.0004	$\hat{g}_A(y)$	1.0241	0.0347
					0.0335
		0.0006	$\hat{g}_M(y)$	1.8264	0.2354
					0.2746

3.5.2 Evaluation of Additive Kernel Density Estimator via Cross-Validation

Now we consider estimating the following target densities by our proposed additive density estimator $\hat{g}_A(y)$:

I. Normal $N(5, 5)$

II. Poisson $Pois(2)$

III. Mixture of two normal $0.5N(1, 1) + 0.5N(5, 1)$

The simulation procedure is consistent with that outlined in Section 3.5.1. We compare the performance of $\hat{g}_A(y)$ using the optimal bandwidth obtained from the cross-validation method to that of $\hat{g}_A(y)$ using the theoretical optimal bandwidth in (3.12), because in reality, we do not have access to the latter, and we can only do bandwidth selection based on the given data.

The Least-Squares (Unbiased) Cross-Validation selector proposed by Scott & Terrell (1987) is defined as

$$\hat{h}_{UCV} := \underset{h>0}{\text{argmin}} \text{UCV}(h)$$

where $\text{UCV}(h) := \int \hat{q}(z; h)^2 dz - (2/n) \sum_{i=1}^{n} \hat{q}_{-i}(Z_i; h)$.

Table 3.4 shows the empirical MISE of the additive kernel density estimator $\hat{g}_A(y)$ when utilizing the above bandwidth selector and the theoretical optimal bandwidth in three distinct target densities. In practice, where the theoretical optimal bandwidth is unknown, the proposed additive kernel density estimator $\hat{g}_A(y)$ performs well via the cross-validation method. The MISE of $\hat{g}_A(y)$ using the UCV bandwidth selector

is fairly close to that of $\hat{g}_A(y)$ using the theoretical optimal bandwidth, and that theoretical value falls within the 95% percentile interval.

Table 3.4. Performance of the additive kernel density estimators with the optimal bandwidth using the cross-validation method and the theoretical value. $T = 1$.

		Original KDE Without Scrambling (T=0)			T=1 Var(S)=0.08	
	n	h_{opt}	MISE		h_{opt}	MISE
Case I: Normal	100	0.9681	0.0024	Cross-Validation	0.9262⁺ (0.3545, 1.1817)*	0.0032
				Theoretical	0.9758	0.0025
	500	0.7016	0.0007	Cross-Validation	0.6680 (0.2582, 0.7947)	0.0009
				Theoretical	0.7072	0.0008
	1000	0.6108	0.0004	Cross-Validation	0.5881 (0.2655, 0.6833)	0.0005
				Theoretical	0.6157	0.0005
Case II: Poisson	100	0.6694	0.0045	Cross-Validation	0.5174 (0.1546, 0.7374)	0.0056
				Theoretical	0.3344	0.0053
	500	0.5006	0.0016	Cross-Validation	0.2214 (0.0964, 0.4549)	0.0019
				Theoretical	0.2424	0.0018
	1000	0.4357	0.0011	Cross-Validation	0.1522 (0.0758, 0.2309)	0.0011
				Theoretical	0.2110	0.0012
Case III: Mixture Normals	100	0.8342	0.0104	Cross-Validation	0.5728 (0.2513, 0.8312)	0.0086
				Theoretical	0.5267	0.0093
	500	0.3883	0.0021	Cross-Validation	0.3749 (0.1900, 0.4752)	0.0036
				Theoretical	0.3818	0.0045
	1000	0.3265	0.0013	Cross-Validation	0.3249 (0.1741, 0.3960)	0.0027
				Theoretical	0.3323	0.0037

+: mean of the bandwidth using the cross-validation method under 1,000 iterations.
*: 2.5th and 97.5th percentiles of the bandwidth using the cross-validation method under 1,000 iterations.

Table 3.4 (Continued). Performance of the additive kernel density estimators with the optimal bandwidth using the cross-validation method and the theoretical value. $T = 5$.

		Original KDE Without Scrambling (T=0)			T=5 Var(S)=2.08	
	n	h_{opt}	MISE		h_{opt}	MISE
Case I: Normal	100	0.9681	0.0024	Cross-Validation	1.0706 (0.3584, 1.3878)	0.0107
				Theoretical	1.1634	0.0109
	500	0.7016	0.0007	Cross-Validation	0.7759 (0.2643, 0.9318)	0.0087
				Theoretical	0.8432	0.0089
	1000	0.6108	0.0004	Cross-Validation	0.6736 (0.2255, 0.7982)	0.0083
				Theoretical	0.7341	0.0084
Case II: Poisson	100	0.6694	0.0045	Cross-Validation	0.8143 (0.2656, 1.0310)	0.0405
				Theoretical	0.6451	0.0389
	500	0.5006	0.0016	Cross-Validation	0.5889 (0.1924, 0.7077)	0.0381
				Theoretical	0.4675	0.0372
	1000	0.4357	0.0011	Cross-Validation	0.4747 (0.1439, 0.6067)	0.0372
				Theoretical	0.4070	0.0368
Case III: Mixture Normals	100	0.8342	0.0104	Cross-Validation	1.0806 (0.3384, 1.3385)	0.0347
				Theoretical	0.8885	0.0328
	500	0.3883	0.0021	Cross-Validation	0.7652 (0.2752, 0.9168)	0.0331
				Theoretical	0.6440	0.0322
	1000	0.3265	0.0013	Cross-Validation	0.6520 (0.2331, 0.7895)	0.0327
				Theoretical	0.5606	0.0321

Table 3.4 (Continued). Performance of the additive kernel density estimators with the optimal bandwidth using the cross-validation method and the theoretical value. $T = 10$.

		Original KDE Without Scrambling (T=0)			T=10 Var(S)=8.33	
	n	h_{opt}	MISE		h_{opt}	MISE
Case I: Normal	100	0.9681	0.0024	Cross-Validation	1.4659 (0.5131, 1.8555)	0.0364
				Theoretical	1.6232	0.0368
	500	0.7016	0.0007	Cross-Validation	1.0919 (0.4419, 1.2673)	0.0350
				Theoretical	1.1764	0.0351
	1000	0.6108	0.0004	Cross-Validation	0.9605 (0.3664, 1.0870)	0.0346
				Theoretical	1.0241	0.0347
Case II: Poisson	100	0.6694	0.0045	Cross-Validation	1.2729 (0.4422, 1.5999)	0.0928
				Theoretical	1.1018	0.0918
	500	0.5006	0.0016	Cross-Validation	0.9138 (0.3531, 1.0999)	0.0908
				Theoretical	0.7985	0.0903
	1000	0.4357	0.0011	Cross-Validation	0.7916 (0.3122, 0.9455)	0.0903
				Theoretical	0.6952	0.0899
Case III: Mixture Normals	100	0.8342	0.0104	Cross-Validation	1.4730 (0.5316, 1.8237)	0.0576
				Theoretical	1.2613	0.0554
	500	0.3883	0.0021	Cross-Validation	1.0564 (0.4200, 1.2582)	0.0562
				Theoretical	0.9142	0.0552
	1000	0.3265	0.0013	Cross-Validation	0.9179 (0.4109, 1.0786)	0.0558
				Theoretical	0.7958	0.0552

3.6 A Numerical Example

In this section, we present a detailed numerical example to illustrate the practical application of the proposed additive kernel density estimator. Consider a hypothetical scenario where we aim to investigate a finite population comprising $N = 5,000$ individuals, with the sensitive variable Y following a normal distribution with a mean of $\mu_Y = 5$ and a variance of $\sigma_Y^2 = 5$. To facilitate our analysis, we draw a single sample of size $n = 500$ from this population and employ the additive RRT model to collect the data. Under this model, the reported response Z is given by $Z = Y + S$, where the scrambling variable S follows a uniform distribution, $U(0, T = 5)$.

Using the UCV bandwidth selector, we obtain the optimal bandwidth h_{opt} as $h_{opt} = 0.5902$ and then plug it into the KDE process to derive the estimation results. The estimation results are presented in Figure 3.1. The MISE of the additive kernel density estimate $\hat{g}_A(y)$ is about 0.0085, which is consistent with the corresponding theoretical AMSE and empirical MISE in Tables 3.3 and 3.4. The additive kernel density estimator can make a reasonably good estimate of the true density curve.

Upon visual inspection, we observe that the additive kernel density estimator closely approximates the true density curve, demonstrating its efficacy in capturing the underlying distribution of the sensitive variable. As depicted in Figure 3.1, both the additive kernel density estimate and the true density curve exhibit similar characteristics, including a bell-shaped distribution that is symmetric and centered around the mean value of 5. While there may be slight deviations in the tails of the true density curve, overall, the additive kernel density estimator provides a reasonably accurate representation of the underlying distribution.

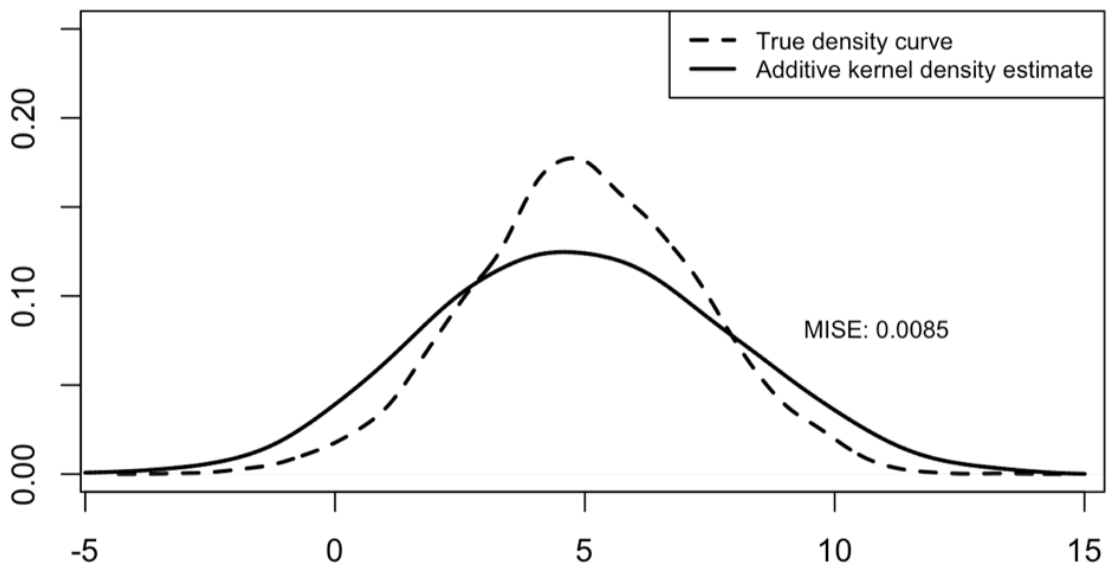

Figure 3.1. Estimation results for the quantitative sensitive variable based on randomized data. The red line is the true density curve $g(y)$ and the black line represents the additive kernel density estimate $\hat{g}_A(y)$. The MISE is also reported.

3.7 Concluding Chapter Remarks

In this Chapter, we introduce a novel kernel density estimator rooted in additive RRT models, which hold widespread utility in survey sampling owing to their numerous advantages over multiplicative RRT models. We highlight the fact that multiplicative models are prone to privacy violation whereas there is no such concern with additive models. Also, respondents with limited mathematical proficiency may find an additive model easier to use. Moreover, through a series of simulation experiments, we demonstrate the superior efficiency of employing additive RRT models in comparison to alternative methodologies.

In practical applications, the challenge of determining the optimal bandwidth for kernel density estimation often arises, especially in the absence of theoretical

guidelines. To address this challenge, we leverage a readily available data-driven approach for bandwidth selection, such as the cross-validation method, to derive an optimal bandwidth directly from the observed data. By adopting this pragmatic approach, researchers can effectively leverage the existing methodology of additive kernel density estimation in real-world survey settings, without the need for complex theoretical calculations.

It is noteworthy that the current literature lacks extensive exploration of the comprehensive distribution of sensitive variables through simulation studies. By integrating such empirical investigations into our study, we not only validate the theoretical foundations of our proposed methodology but also contribute to a broader understanding of this burgeoning field of inquiry. We aspire that our research will provide fresh perspectives and insights, enriching the discourse surrounding the estimation of sensitive variables in survey research and fostering further exploration in this evolving domain.

Chapter 4: Kernel Density Estimation of a Sensitive Variable in the Presence of Auxiliary Information

4.1 Introduction

In this Chapter, we delve deeper into the research conducted by Shou & Gupta (2023) (see Chapter 3), aiming to broaden its scope and implications. Our focus shifts towards the integration of auxiliary variables within the RRT framework, a novel approach that has the potential to refine the precision of density estimators and unveil novel insights into the distribution patterns of sensitive variables under RRT models.

Drawing inspiration from the work of Mostafa & Ahmad (2019), we propose a novel methodology that augments the traditional kernel density estimator with auxiliary variables, thereby enriching the accuracy and robustness of the density estimation process. By leveraging auxiliary information in tandem with the core RRT framework, we aim to unlock new dimensions of understanding regarding the distributional characteristics of sensitive variables.

To empirically validate the efficacy of our proposed methodology, we conduct extensive simulations under varied conditions and scenarios. These simulations serve as a comprehensive evaluation of the proposed approach, shedding light on its performance across a spectrum of factors including noise levels, sample size, and the correlation between auxiliary variables and sensitive variables.

Through this rigorous empirical analysis, we aim to elucidate the advantages of incorporating auxiliary information into the RRT framework, demonstrating its potential to enhance the accuracy and reliability of density estimators in sensitive survey research.

4.2 Proposed Kernel Density Estimator

We extend the additive kernel density estimator presented by Shou & Gupta (2023) (see Chapter 3) by incorporating auxiliary information within the RRT framework, drawing inspiration from the methodology developed by Mostafa & Ahmad (2019) (see Section 2.2.1).

Let Y be the sensitive study variable with mean μ_Y and variance σ_Y^2. Let X be a non-sensitive auxiliary variable with mean μ_X and variance σ_X^2, which satisfies the condition in (4.1). Let S be a scrambling variable with mean μ_S and variance σ_S^2, independent of Y, and S follows a uniform distribution of $U(0, T)$, where T is a pre-selected number. The respondent is asked to report a scrambled response for Y given by $Z = Y + S$ in the context of additive RRT models.

Consider a finite population N where the scrambled response Z is observed in n units and is predictable in $(N - n)$ units. Assume that the relationship between Z and the auxiliary variable X can be described by the following parametric regression

model:

$$z_i = \mu(x_i, \boldsymbol{\beta}) + \sigma(x_i)\epsilon_i; i \in U, \tag{4.1}$$

where $\mu(\cdot,\cdot)$ is a known mean function, $\sigma(\cdot,\cdot)$ is a known, strictly positive function, $\boldsymbol{\beta}$ is the unknown model parameter vector, and $\epsilon_i's$ are independent and identically distributed (i.i.d.) random variables with zero mean and unit variance.

Under simple random sampling (SRS), the proposed additive kernel density estimator with auxiliary information is given by

$$\hat{g}_{Aux}(y) = \frac{1}{T}\int_y^{y+T} \hat{q}^*_{kde}(w)dw, \tag{4.2}$$

where

$$\hat{q}^*_{kde}(z) = \frac{1}{n}\sum_{i\in s}\{K_h(z - z_i) - K_h(z - \hat{\beta}x_i)\} + \frac{1}{N}\sum_{i\in U} K_h(z - \hat{\beta}x_i) \tag{4.3}$$

with the kernel $K(x)$ and the bandwidth h.

4.3 Efficiency and Bandwidth Selection for the Proposed Kernel Density Estimator

To obtain the asymptotic mean integrated squared error (AMISE) of the proposed kernel density estimator $\hat{g}_{Aux}(y)$, we first need to find the asymptotic mean squared error (AMSE) of $\hat{q}^*_{kde}(z)$ as follows:

$$Bias(\hat{q}^*_{kde}(z)) = \frac{1}{2}h^2 c_K q''(z) + O(h^2), \tag{4.4}$$

where $c_K = \int y^2 K(y)dy$.

$$\begin{aligned} Var(\hat{q}^*_{kde}(z)) &= \frac{\delta}{nh^3}\big[\beta^2 E(X^2) - 2\beta E(XZ) + E(Z^2)\big]\{K'(z)\}^2 + O(\frac{1}{Nh^3}) \\ &= \frac{\delta}{nh^3}\big[\beta^2(\sigma_X^2 + \mu_X^2) - 2\beta(\beta \cdot \sigma_X^2 + \mu_X(\mu_Y + \mu_S)) + \sigma_Y^2 + \sigma_S^2 + (\mu_Y \\ &\quad + \mu_S)^2\big]\{K'(z)\}^2 + O(\frac{1}{Nh^3}), \end{aligned} \tag{4.5}$$

where $\delta = 1 - n/N$.

Then

$$\begin{aligned} \text{MSE}(\hat{q}^*_{kde}(z)) &= Bias^2(\hat{q}_{KDE^*}(z)) + Var(\hat{q}_{KDE^*}(z)) \\ &= \frac{1}{4}h^4 c_K^2\{q''(z)\}^2 + \frac{\delta}{nh^3}\big[(\beta\mu_X - \mu_Y)^2 - \beta^2\sigma_X^2 - 2\beta\mu_X\mu_S + \sigma_Y^2 + \sigma_S^2 \\ &\quad + \mu_S^2 + 2\mu_Y\mu_S\big] + O(h^4 + \frac{1}{Nh^3}) \end{aligned} \tag{4.6}$$

and

$$\begin{aligned} \text{AMSE}(\hat{q}^*_{kde}(z)) \approx \frac{1}{4}h^4 c_K^2\{q''(z)\}^2 + \frac{\delta}{nh^3}\big[(\beta\mu_X - \mu_Y)^2 - \beta^2\sigma_X^2 - 2\beta\mu_X\mu_S + \sigma_Y^2 + \sigma_S^2 \\ + \mu_S^2 + 2\mu_Y\mu_S\big]. \end{aligned} \tag{4.7}$$

Now we have

$$\begin{aligned} \text{MSE}(\hat{g}_{Aux}(y)) &= E\bigg(\hat{g}_{Aux}(y) - g(y)\bigg)^2 \\ &= \frac{1}{T^2}\int_y^{y+T} \text{MSE}(\hat{q}^*_{kde}(w))dw + \frac{1}{T^2}\Big(\int_y^{y+T} q(w)dw\Big)^2 + \Big(g(y)\Big)^2 \\ &\quad - \frac{2g(y)}{T}\int_y^{y+T} q(w)dw. \end{aligned} \tag{4.8}$$

From (4.7) and (4.8), using the Taylor's approximation and retaining terms of order up to 2, the AMSE and AMISE of $\hat{g}_{Aux}(y)$ are given by

$$\begin{aligned}\text{AMSE}(\hat{g}_{Aux}(y)) \approx \frac{h^4 c_K^2}{4T^2}\int_y^{y+T}\big(q''(w)\big)^2 dw + \frac{\delta M}{nh^3T}\{K'(y)\}^2 \\ + \frac{1}{T^2}\Big(\int_y^{y+T} q(w)dw\Big)^2 + \big(g(y)\big)^2 \\ - \frac{2g(y)}{T}\int_y^{y+T} q(w)dw \end{aligned} \tag{4.9}$$

and

$$\begin{aligned}\text{AMISE}(\hat{g}_{Aux}(y)) \approx \frac{h^4 c_K^2}{4T^2}\iint_y^{y+T}\big(q''(w)\big)^2 dwdy + \frac{\delta M}{nh^3T}\int\{K'(y)\}^2 dy \\ + \frac{1}{T^2}\int\Big(\int_y^{y+T} q(w)dw\Big)^2 dy + \int\big(g(y)\big)^2 dy \\ - \frac{2}{T}\int g(y)\int_y^{y+T} q(w)dwdy, \end{aligned} \tag{4.10}$$

where $M = (\beta\mu_X - \mu_Y)^2 - \beta^2\sigma_X^2 - 2\beta\mu_X\mu_S + \sigma_Y^2 + \sigma_S^2 + \mu_S^2 + 2\mu_Y\mu_S$.

Differentiating (4.10) with respect to h, we get the following optimum value:

$$h_{opt,Aux} = \left(\frac{3\delta MT\int\{K'(y)\}^2 dy}{nc_K^2\iint_y^{y+T}\big(q''(w)\big)^2 dwdy}\right)^{1/7}, \tag{4.11}$$

which minimizes the AMISE.

4.4 Simulation Study

In this section, we present the results of a simulation study, with a specific emphasis on evaluating the performance of the proposed kernel density estimator $\hat{g}_{Aux}(y)$ when

utilizing a non-sensitive auxiliary variable, in contrast to the additive kernel density estimator $\hat{g}_A(y)$ discussed in Section 3, where auxiliary information is ignored.

4.4.1 Simulation Procedure

In the simulation study, we consider a finite population of size $N = 5,000$ generated from a normal distribution with mean $\mu_Y = 5$ and variance $\sigma_Y^2 = 5$. The scrambling variable S is taken to be a uniform variate from $U(0, T)$, where T can take on values of 1, 5, and 10. The reported response is given by $Z = Y + S$ under additive RRT models.

Inspired by Mostafa & Ahmad (2019), three models, denoted as Models I–III, are utilized in the simulation study to represent three distinct forms of the relationship between X and Y. These models include a linear relationship (Model I: $y = 1 + 2(x - 0.5) + \epsilon$), a logarithmic relationship (Model II: $y = 2.5 \log(x + 1.5) + \epsilon$), and a hard-to-detect nonlinear relationship (Model III: $y = \pm\sqrt{2x} + 0.6\epsilon$), respectively. In each model, the errors ϵ are regulated to achieve both low ($\rho \approx 0.36$) and high ($\rho \approx 0.75$) correlations between X and Y.

We consider three sample sizes: $n = \{100, 500, 1000\}$, using SRSWOR (simple random sampling without replacement). We choose the normal kernel when using KDE, which means $K(x) = \phi(x)$, and ϕ is the standard normal density function. The bandwidth h is determined through the Least-Squares (Unbiased) Cross-Validation (UCV) selector, which is defined as

$$\hat{h}_{UCV} := \underset{h>0}{\text{argmin}} \text{UCV}(h),$$

where $\text{UCV}(h) := \int \hat{q}(z; h)^2 dz - (2/n) \sum_{i=1}^{n} \hat{q}_{-i}(Z_i; h)$.

Coding for the simulations was done in R and results are averaged over 500 iterations. The empirical MISE of the kernel density estimator $\hat{g}(y)$ is computed by

$$\text{MISE}_E(\hat{g}(y)) = \frac{1}{500}\sum_{i=1}^{500}\int \big(\hat{g}(y) - g(y)\big)^2 dy,$$

where $\hat{g}(y) = \{\hat{g}_A(y), \hat{g}_{Aux}(y)\}$.

4.4.2 Simulation Results

The results from the simulations, as detailed in Table 4.1, illustrate how the performance of the proposed kernel density estimator $\hat{g}_{Aux}(y)$, which incorporates the auxiliary variable X, compares to that of the additive kernel density estimator $\hat{g}_A(y)$, which does not use X, across different scenarios. For higher values of T ($T = 5$ and 10), i.e. with greater noise, it is evident that the additive kernel density estimator $\hat{g}_A(y)$ not utilizing the auxiliary variable X yields the highest MISE since it does not leverage the auxiliary information. In contrast, the proposed kernel density estimator $\hat{g}_{Aux}(y)$ using the auxiliary variable X with a high correlation to the study variable Y yields the smallest MISE, demonstrating the advantage of incorporating relevant auxiliary information. Additionally, the proposed kernel density estimator $\hat{g}_{Aux}(y)$ employing the auxiliary variable X with a low correlation to the study variable Y produces a worse MISE compared to the high correlation case, yet still outperforms the additive kernel density estimator $\hat{g}_A(y)$ not using X.

We may also note that when $T = 1$, utilizing the auxiliary variable X results in a worse MISE compared to not utilizing it. This observation can be attributed to the relatively small amount of noise present at $T = 1$, which has minimal impact on the estimation process. In other words, when using the approximation in (4.4), (4.5), and

(4.6), some inaccuracies emerge. These inaccuracies become more pronounced when there is a low degree of noise ($T = 1$) but tend to be less noticeable when the degree of noise is higher ($T = 5$ and 10). As evidenced by the second and third columns in Table 4.1, the MISE values with scrambling are very similar to the original MISE values without scrambling. Consequently, when we employ auxiliary information to predict non-sampled z's and subsequently increase the sample size, more variation is introduced into the estimation process, leading to an eventual increase in MISE, particularly when the correlation between X and Y is low.

Table 4.1. Theoretical (**bold**) AMISEs and empirical MISEs of the proposed kernel density estimator $\hat{g}_{Aux}(y)$ with auxiliary variable X and the additive kernel density estimator $\hat{g}_A(y)$ without X. High correlation ($\rho \approx 0.75$) with X.

		T=0 (Without	T=1 Var(S)=0.08			T=5 Var(S)=2.08			T=10 Var(S)=8.33		
ρ	n	Scrambling)	$\hat{g}_A(y)$	(Model)	$\hat{g}_{Aux}(y)$	$\hat{g}_A(y)$	(Model)	$\hat{g}_{Aux}(y)$	$\hat{g}_A(y)$	(Model)	$\hat{g}_{Aux}(y)$
0.75	100	.0024	.0032 **.0025**	I	.0065 **.0051**	.0107 **.0109**	I	.0009 **.0008**	.0364 **.0368**	I	.0206 **.0202**
				II	.1103 **.1063**		II	.0067 **.0065**		II	.0227 **.0222**
				III	.0425 **.0384**		III	.0029 **.0027**		III	.0215 **.0211**
	500	.0007	.0009 **.0008**	I	.0052 **.0046**	.0087 **.0089**	I	.0004 **.0004**	.0350 **.0351**	I	.0209 **.0207**
				II	.1060 **.0942**		II	.0055 **.0054**		II	.0228 **.0225**
				III	.0379 **.0337**		III	.0022 **.0022**		III	.0216 **.0213**
	1000	.0004	.0005 **.0005**	I	.0042 **.0039**	.0083 **.0084**	I	.0005 **.0005**	.0346 **.0347**	I	.0218 **.0217**
				II	.0866 **.0801**		II	.0038 **.0037**		II	.0232 **.0231**
				III	.0298 **.0276**		III	.0017 **.0017**		III	.0222 **.0221**

Table 4.1 (Continued). Theoretical (**bold**) AMISEs and empirical MISEs of the proposed kernel density estimator $\hat{g}_{Aux}(y)$ with auxiliary variable X and the additive kernel density estimator $\hat{g}_A(y)$ without X. Low correlation ($\rho \approx 0.36$) with X.

		T=0	T=1			T=5			T=10		
		(Without	Var(S)=0.08			Var(S)=2.08			Var(S)=8.33		
ρ	n	Scrambling)	$\hat{g}_A(y)$	(Model)	$\hat{g}_{Aux}(y)$	$\hat{g}_A(y)$	(Model)	$\hat{g}_{Aux}(y)$	$\hat{g}_A(y)$	(Model)	$\hat{g}_{Aux}(y)$
0.36	100	.0024	.0032	I	.0976	.0107	I	.0072	.0364	I	.0220
			.0025		**.0929**	**.0109**		**.0070**	**.0368**		**.0217**
				II	.2135		II	.0155		II	.0269
					.1679			**.0152**			**.0263**
				III	.1470		III	.0099		III	.0251
					.1563			**.0096**			**.0243**
	500	.0007	.0009	I	.0942	.0087	I	.0053	.0350	I	.0219
			.0008		**.0887**	**.0089**		**.0052**	**.0351**		**.0216**
				II	.3359		II	.0126		II	.0273
					.2986			**.0123**			**.0271**
				III	.1769		III	.0072		III	.0245
					.1472			**.0070**			**.0241**
	1000	.0004	.0005	I	.0737	.0083	I	.0030	.0346	I	.0222
			.0005		**.0682**	**.0084**		**.0030**	**.0347**		**.0221**
				II	.2691		II	.0106		II	.0245
					.2490			**.0104**			**.0244**
				III	.1528		III	.0053		III	.0228
					.1384			**.0052**			**.0227**

4.5 Concluding Chapter Remarks

In this Chapter, we proposed a kernel density estimator under additive RRT models in the presence of auxiliary information. Our study explores the integration of a non-sensitive auxiliary variable to enhance the accuracy of estimating the distribution of a sensitive variable. This innovative approach aims to maintain respondent privacy while improving the precision of sensitive data analysis. Drawing inspiration from prior research, we present our methodology and conduct a comprehensive simulation study to assess its performance. The results shed light on the influence of noise, sample size, and the correlation between the study variable and auxiliary variable on the accuracy of estimation, underscoring the advantages of incorporating auxiliary data within additive RRT models.

Chapter 5: Kernel Density Estimation Using Optional Randomized Response Technique Models

5.1 Introduction

In this Chapter, we introduce an extension to the kernel density estimator within the framework of KDE using additive RRT models. Our aim is to broaden the scope of direct distribution estimation by accommodating scenarios where respondents have the flexibility to choose between providing true or scrambled responses, facilitated by the incorporation of sensitivity level denoted as W. This sensitivity level W has been extensively studied in the context of population parameter estimation using RRT models. These include Gupta et al. (2014), Kalucha et al. (2016), Gupta et al. (2017), Khalil et al. (2021), Mehta & Aggarwal (2018), and Narjis & Shabbir (2020).

Our research involves deriving the theoretical results for this extended estimator and conducting a comprehensive simulation study to assess its performance under various conditions. By incorporating the sensitivity level W, the proposed estimator

holds the potential to provide higher accuracy and robustness in handling sensitive data scenarios, making it a crucial contribution to the field of direct distribution estimation using RRT models.

5.2 Proposed Kernel Density Estimator

We propose a kernel density estimator within the framework of optional RRT models using additive scrambling. Let Y denote the sensitive study variable with a mean of μ_Y and a variance of σ_Y^2. Let S represent a scrambling variable with a mean of μ_S and a variance of σ_S^2, which is independent of Y. S follows a uniform distribution denoted by $U(0,T)$, where T is a pre-selected number. Let W be the sensitivity level of the underlying sensitive question. In this model, the reported response Z is given by

$$Z = \begin{cases} Y, & \text{with probability } 1-W \\ Y+S, & \text{with probability } W \end{cases} \tag{5.1}$$

Let $F(s), G(y), Q(z)$ denote the CDF of S, Y, Z, respectively, with corresponding pdf $f(s), g(y)$, and $q(z)$.

Since $S \sim U(0,T)$, $F(s) = (1-W) + sW/T, s \in [0,T]$.

$$\begin{aligned} G(y) &= P(Z - S \leq y) \\ &= \int_0^\infty q(z)[1 - F(z-y)]dz \\ &= W\int_y^{y+T} \left(1 - \frac{z-y}{T}\right) dQ(z) + Wq(y) + Q(y) - Q(0) \\ &= WQ(y+T) + (1-W)Q(y) - \frac{W}{T}\int_y^{y+T} (z-y)q(z)dz + Wq(y). \end{aligned} \tag{5.2}$$

Taking the derivative of both sides of (5.2), we get

$$\begin{aligned} g(y) &= Wq(y+T) + (1-W)q(y) - \frac{W}{T}\Big((y+T)q(y+T) - yq(y)\Big) \\ &\quad + \frac{W}{T}\int_y^{y+T} q(z)dz + \frac{Wy}{T}\Big(q(y+T) - q(y)\Big) + Wq'(y) \\ &= (1-W)q(y) + \frac{W}{T}\int_y^{y+T} q(z)dz + Wq'(y) \end{aligned} \tag{5.3}$$

Using KDE, $g(y)$ can be estimated by

$$\hat{g}(y) = (1-W)\hat{q}_{KDE}(z) + \frac{W}{T}\int_y^{y+T} \hat{q}_{KDE}(z)dz + W\hat{q}'_{KDE}(z), \tag{5.4}$$

where $\hat{q}_{KDE}(z) = (1/nh)\sum_{i=1}^{n} K[(z - Z_i)/h]$ with the kernel $K(x)$ and the bandwidth h.

The sensitivity level W in (5.4) can be estimated through a pre-survey utilizing the binary RRT model introduced by Warner (1965) (see Section 2.1.1). This method not only enables the estimation of W but also offers guidance on the preferred utilization of optional RRT models.

From (2.2) and (2.3), an unbiased estimator of W is given by

$$\hat{W} = \frac{\hat{p_y} - (1-p)}{2p-1}, \tag{5.5}$$

where p is a pre-determined parameter denoting the probability of a respondent answering the sensitive question directly during the pre-survey, $p \neq \frac{1}{2}$. $\hat{p_y} = \frac{m_1}{m}$, where m_1 denotes the count of respondents answering 'yes' within a sample of size m.

The variance of the above estimator is

$$Var(\hat{W}) = \frac{W(1-W)}{m} + \frac{p(1-p)}{m(2p-1)^2}. \tag{5.6}$$

We now propose our additive kernel density estimator utilizing optional RRT models, defined as

$$\hat{g}_W(y) = (1-\hat{W})\hat{q}_{KDE}(z) + \frac{\hat{W}}{T}\int_y^{y+T} \hat{q}_{KDE}(z)dz + \hat{W}\hat{q}'_{KDE}(z). \tag{5.7}$$

5.3 Efficiency and Bandwidth Selection in Optional Additive Kernel Density Estimator

To estimate the mean integrated square error (MISE) of the proposed kernel density estimator $\hat{g}_W(y)$, we first compute the mean square error (MSE) of $\hat{g}_W(y)$ at any point y as follows:

$$\begin{aligned}
\text{MSE}(\hat{g}_W(y)) &= E\bigg(\hat{g}_W(y) - g(y)\bigg)^2 \\
&= E\bigg((1-\hat{W})\hat{q}_{KDE}(z) + \frac{\hat{W}}{T}\int_y^{y+T} \hat{q}_{KDE}(z)dz + \hat{W}\hat{q}'_{KDE}(z) - g(y)\bigg)^2 \\
&= E\bigg(W - \hat{W})\hat{q}_{KDE}(y) + \frac{\hat{W}-W}{T}\int_y^{y+T} \hat{q}_{KDE}(z)dz \\
&\quad + \frac{\hat{W}-W}{T}\hat{q}'_{KDE}(y) + (1-W)(\hat{q}_{KDE}(y) - q(y)) \\
&\quad + \frac{W}{T}\int_y^{y+T} (\hat{q}_{KDE}(y) - q(y))dz + W(\hat{q}'_{KDE}(y) - q(y)) \\
&\quad + (1-W)q(y) + \frac{W}{T}\int_y^{y+T} q(z)dz + Wq'(y) - g(y)\bigg)^2
\end{aligned}$$

$$\begin{aligned}\mathrm{MSE}(\hat{g}_W(y)) = q^2(y)\mathrm{MSE}(\hat{W}) + \frac{1}{T^2}\left(\int_y^{y+T} q(z)dz\right)^2 \mathrm{MSE}(\hat{W}) + \frac{1}{T^2}q'^2(y)\mathrm{MSE}(\hat{\hat{W}}) \\ + (1-W)^2\mathrm{MSE}(\hat{q}_{KDE}(y)) + \frac{W^2}{T^2}\int_y^{y+T} \mathrm{MSE}(\hat{q}_{KDE}(z))dz \\ + W^2\mathrm{MSE}(\hat{q}'_{KDE}(y)) + \Bigg((1-W)q(y) + \frac{W}{T}\int_y^{y+T} q(z)dz \\ + Wq'(y) - g(y)\Bigg)^2 \end{aligned} \tag{5.8}$$

Wand & Jones (1994) provided an expression for the approximate MSE (AMSE) in the context of kernel density estimation, offering a generalized framework. Thus, for Model (5.1), the AMSE of $\hat{q}_{KDE}(u)$ can be represented as follows:

$$\mathrm{AMSE}(\hat{q}_{KDE}(u)) = \frac{h^4 c_K^2 \big(q''(u)\big)^2}{4} + \frac{q(u)}{nh}\int \big(K(y)\big)^2 dy, \tag{5.9}$$

where $c_K = \int y^2 K(y)dy$.

Then substituting (5.7) and (5.9), using the Taylor's approximation, and retaining terms of order up to 2, the approximate mean integrated square error (AMISE) of $\hat{g}_W(y)$ is given by

$$\mathrm{AMISE}(\hat{g}_W(y)) = \int \mathrm{AMSE}(\hat{g}_W(y))dy$$

$$
\begin{aligned}
\text{AMISE}(\hat{g}_W(y)) = &\left(\frac{W(1-W)}{m} + \frac{p(1-p)}{m(2p-1)^2}\right)\left(\int q^2(y)dy\right. \\
&+ \frac{1}{T^2}\int \Big(\int_y^{y+T} q(z)dz\Big)^2 dy + \left.\frac{1}{T^2}\int q'^2(y)dy\right) \\
&+ (1-W)^2\left(\frac{h^4 c_K^2}{4}\int \big(q''(y)\big)^2 dy + \frac{\int \big(K(y)\big)^2 dy)}{nh}\right) \\
&+ W^2\left(\frac{h^4 c_K^2}{4T^2}\iint_y^{y+T} \big(q''(z)\big)^2 dzdy + \frac{\int \big(K(y)\big)^2 dy}{nhT^2}\iint_y^{y+T} q(z)dzdy\right) \\
&+ \int\left((1-W)q(y) + \frac{W}{T}\int_y^{y+T} q(z)dz + Wq'(y) - g(y)\right)^2 dy
\end{aligned}
\tag{5.10}
$$

By differentiating (5.10) with respect to h, we derive the optimal value of h as follows:

$$
h_{\text{opt},W} = \left(\frac{\frac{\int (K(y))^2 dy}{n}\big((1-W)^2 + \frac{W^2}{T^2}\iint_Y^{y+T} q(z)dz\big)}{c_K^2\big((1-W)^2\int (q''(y))^2 dy + \frac{W^2}{T^2}\iint_y^{y+T}(q''(z))^2 dz\big)}\right)^{1/5}, \tag{5.11}
$$

which minimizes the AMISE.

5.4 Simulation Study

We conduct a simulation study to assess the performance of our proposed additive kernel density estimator $\hat{g}_W(y)$ utilizing an optional RRT model, in contrast to the original additive kernel density estimator $\hat{g}_A(y)$ employing a non-optional RRT model.

In the simulation study, we consider a finite population of size $N = 10,000$, generated from a normal distribution with a mean of $\mu_Y = 5$ and a variance of $\sigma_Y^2 = 5$. The scrambling variable S is uniformly distributed from $U(0, T)$, where $T = 5, 10$. For

the non-optional additive RRT model, the reported response is given by $Z_A = Y + S$; for the optional additive RRT model, the reported response is defined in Model (5.1).

We consider three sample sizes: $n = \{100, 500, 1000\}$, employing simple random sampling without replacement (SRSWOR). For KDE, we opt for the normal kernel, denoted as $K(x) = \phi(x)$, where ϕ represents the standard normal density function. The bandwidth h is determined using the Least-Squares (Unbiased) Cross-Validation (UCV) selector, defined as:

$$\hat{h}_{UCV} := \underset{h>0}{\text{argmin}} \text{UCV}(h),$$

where $\text{UCV}(h) := \int \hat{q}(z;h)^2 dz - (2/n) \sum_{i=1}^{n} \hat{q}_{-i}(Z_i;h)$.

Coding for the simulations was done in R and results are averaged over 1,000 iterations. The empirical MISE of the kernel density estimator $\hat{g}(y)$ is computed by

$$\text{MISE}_E(\hat{g}(y)) = \frac{1}{1000} \sum_{i=1}^{1000} \int \big(\hat{g}(y) - g(y)\big)^2 dy,$$

where $\hat{g}(y) = \{\hat{g}_A(y), \hat{g}_W(y)\}$.

In Table 5.1, rows with $W = 1$ display theoretical (**bold**) AMISEs and empirical MISEs of the additive kernel density estimator $\hat{g}_A(y)$ under a non-optional RRT model, while the remaining rows with $W = 0.1, ..., 0.99$ display theoretical (**bold**) AMISEs and empirical MISEs of our proposed kernel density estimator $\hat{g}_W(y)$ under an optional RRT model. It is evident that the theoretical AMISEs closely match the empirical MISEs for both kernel density estimators $\hat{g}_A(y)$ and $\hat{g}_W(y)$ using different RRT models.

Observations from Table 5.1 suggest that, for high noise levels ($T = 5$ and 10), the proposed kernel density estimator $\hat{g}_W(y)$ under an optional RRT model tends to

outperform the additive kernel density estimator $\hat{g}_A(y)$ under a non-optional RRT model in most cases. As n increases, the performance of $\hat{g}_W(y)$ improves significantly. For instance, when $W = 0.2$, the empirical MISE of $\hat{g}_W(y)$ decreases by 12.5% for $n = 100$, 36.2% for $n = 500$, and 41.1% for $n = 1000$ upon incorporating optionality into additive RRT models.

Moreover, as T increases, the performance of $\hat{g}_W(y)$ also improves noticeably. For the same $W = 0.2$, the empirical MISE of $\hat{g}_W(y)$ decreases by 62.3% for $n = 100$, 69.0% for $n = 500$, and 70.9% for $n = 1000$ when optionality is incorporated into additive RRT models.

We may also note that when W approaches 0.5, using optional RRT models does not guarantee improved performance. This observation may arise from the increased ambiguity and complexity introduced when respondents are equally likely to provide either true or scrambled responses. In such cases, accurately estimating the sensitivity level and effectively integrating scrambled responses into the analysis becomes challenging.

Table 5.1. Theoretical (**bold**) AMISEs and empirical MISEs of the proposed kernel density estimator $\hat{g}_W(y)$ with optional RRT model and the additive kernel density estimator $\hat{g}_A(y)$ with non-optional RRT model. $n = 100$.

	T=0 (Without Scrambling)		T=5	T=10
			Var(S)=2.08	Var(S)=8.33
	Theoretical AMISE	W	Empirical MISE/ Theoretical AMISE	Empirical MISE/ Theoretical AMISE
n=100	**0.00364**	1 (non-optional)	0.00865 **0.00794**	0.03684 **0.03372**
		0.99	0.00610 **0.00516**	0.03144 **0.02583**
		0.9	0.00697 **0.00615**	0.03511 **0.03007**
		0.8	0.00861 **0.00769**	0.03815 **0.03373**
		0.7	0.01050 **0.00966**	0.03943 **0.03565**
		0.6	0.01132 **0.01023**	0.03804 **0.03435**
		0.5	0.01129 **0.01035**	0.03442 **0.03071**
		0.4	0.01043 **0.00936**	0.02884 **0.02499**
		0.3	0.00917 **0.00786**	0.02137 **0.01809**
		0.2	0.00757 **0.00643**	0.01390 **0.01154**
		0.1	0.00578 **0.00521**	0.00763 **0.00662**

Table 5.1 (Continued). Theoretical (**bold**) AMISEs and empirical MISEs of the proposed kernel density estimator $\hat{g}_W(y)$ with optional RRT model and the additive kernel density estimator $\hat{g}_A(y)$ with non-optional RRT model. $n = 500$.

	T=0 (Without Scrambling)		T=5	T=10
			Var(S)=2.08	Var(S)=8.33
	Theoretical AMISE	W	Empirical MISE/ Theoretical AMISE	Empirical MISE/ Theoretical AMISE
n=500	**0.00101**	1 (non-optional)	0.00782 **0.00751**	0.03514 **0.03356**
		0.99	0.00501 **0.00463**	0.02824 **0.02564**
		0.9	0.00601 **0.00569**	0.03207 **0.02989**
		0.8	0.00760 **0.00725**	0.03526 **0.03354**
		0.7	0.00927 **0.00915**	0.03639 **0.03539**
		0.6	0.01001 **0.00971**	0.03519 **0.03397**
		0.5	0.00972 **0.00952**	0.03128 **0.03012**
		0.4	0.00873 **0.00829**	0.02545 **0.02413**
		0.3	0.00686 **0.00641**	0.01826 **0.01683**
		0.2	0.00499 **0.00455**	0.01089 **0.00977**
		0.1	0.00290 **0.00276**	0.00452 **0.00426**

Table 5.1 (Continued). Theoretical (**bold**) AMISEs and empirical MISEs of the proposed kernel density estimator $\hat{g}_W(y)$ with optional RRT model and the additive kernel density estimator $\hat{g}_A(y)$ with non-optional RRT model. $n = 1000$.

	T=0 (Without Scrambling)		T=5	T=10
			Var(S)=2.08	Var(S)=8.33
	Theoretical AMISE	W	Empirical MISE/ Theoretical AMISE	Empirical MISE/ Theoretical AMISE
n=1000	**0.00058**	1 (non-optional)	0.00767	0.03472
			0.00743	**0.03354**
		0.99	0.00479	0.02702
			0.00454	**0.02561**
		0.9	0.00582	0.03121
			0.00562	**0.02986**
		0.8	0.00738	0.03461
			0.00718	**0.03351**
		0.7	0.00914	0.03568
			0.00907	**0.03535**
		0.6	0.00972	0.03442
			0.00962	**0.03391**
		0.5	0.00955	0.03066
			0.00938	**0.03003**
		0.4	0.00836	0.02483
			0.00811	**0.02399**
		0.3	0.00653	0.01763
			0.00618	**0.01662**
		0.2	0.00451	0.01010
			0.00425	**0.00949**
		0.1	0.00250	0.00410
			0.00237	**0.00387**

5.5 Concluding Chapter Remarks

In this Chapter, we proposed and evaluated an additive kernel density estimator under an optional RRT model, comparing its performance to the original additive kernel density estimator under a non-optional RRT model. Our findings reveal that the proposed additive kernel density estimator generally outperforms the original additive kernel density estimator, especially in scenarios with high noise levels, indicating the enhanced accuracy and robustness of density estimation when optionality is incorporated into RRT models. Moreover, increasing the sample size leads to significant improvements in the performance of proposed additive kernel density estimator. Higher noise levels also contribute to improved performance. However, challenges arise when the sensitivity level approaches 0.5, diminishing the benefits of optional RRT models due to increased ambiguity in distinguishing between true and scrambled responses. Overall, our study highlights the potential of optional RRT models in enhancing the accuracy and reliability of kernel density estimation, particularly in sensitive data scenarios. By allowing respondents the flexibility to choose between providing true or scrambled responses, optional RRT models offer a valuable tool for mitigating respondent bias and enhancing data quality in scenarios where respondents may be reluctant to disclose sensitive information.

Chapter 6: Concluding Remarks and Future Directions

6.1 Concluding Remarks

In this dissertation, we addressed the challenge of estimating distributions of sensitive variables by investigating kernel density estimation (KDE) under various randomized response technique (RRT) models. The research focused on refining prior methodologies and exploring new avenues to improve the accuracy and efficiency of distribution estimation in sensitive data scenarios.

We introduced a kernel density estimator based on additive RRT models, leveraging their widespread use and advantageous properties in survey sampling. Through comprehensive simulation studies, it was also demonstrated that employing additive RRT models leads to significant improvements in the efficiency of our kernel density estimator. Furthermore, the incorporation of auxiliary information proved instrumental in enhancing the precision of sensitive variable distribution estimation, showcasing the potential of additive RRT models in integrating non-sensitive auxiliary variables.

Moreover, the investigation extended to include optional RRT models, which offer respondents the flexibility to choose between providing true or scrambled responses.

The findings indicated that optional RRT models enhance the accuracy and reliability of KDE, particularly in scenarios with high noise levels.

Overall, the research contributes to advancing estimation techniques for sensitive variables, offering valuable insights into their distribution while maintaining respondent privacy. By combining the strengths of KDE with various RRT models, the study provides a robust framework for addressing the challenges associated with sensitive data analysis in survey sampling. It is hoped that these findings will stimulate further research in this growing area and facilitate the development of more effective methodologies for handling sensitive data in surveys.

6.2 Future Directions

One potential area of focus involves investigating intermediate sensitivity levels within optional RRT models. As observed in our study, challenges arise when sensitivity levels approach 0.5, diminishing the benefits of optionality and introducing ambiguity in response interpretation. Future research could delve into strategies for mitigating the impact of intermediate sensitivity levels on density estimation, exploring innovative approaches to improve the accuracy and reliability of estimation under such conditions.

Another promising direction is the integration of auxiliary information and optionality within additive RRT models to enhance the efficiency of kernel density estimators. Additionally, exploring a generalized kernel density estimator under optionality, which includes both additive and multiplicative scrambling, could optimize the balance between privacy protection and survey efficiency.

In the framework where both additive and multiplicative scrambling are used, respondents who consider a question as sensitive and trust the models could use

additive scrambling for their true responses. Conversely, respondents who view the question as sensitive but do not trust the models could apply a combination of additive and multiplicative scrambling. This dual-layer approach could optimize the balance between privacy protection and survey efficiency, potentially making respondents feel more comfortable providing their answers to sensitive survey questions.

Moreover, there is significant potential in conducting a thorough comparison between mean estimates derived from the proposed kernel density estimators and those obtained through population parameter estimation within the framework of additive RRT models. Inspired by earlier simulation results, a meticulous analysis of the performance of direct distribution estimation compared to traditional population parameter estimation offers a promising avenue for extracting valuable insights. By undertaking comprehensive evaluations, researchers can discern the strengths and limitations of each approach, leading to a greater understanding of their efficacy across varied scenarios.

Additionally, applying our proposed methods to real-world data can provide further validation. This would involve selecting response data for sensitive questions collected using an anonymous survey method, scrambling it, and evaluating the estimators' performance in recovering the distribution of true non-scrambled responses. Such practical applications can help to confirm the robustness and reliability of the proposed estimators in real-world settings, offering a more comprehensive perspective on their effectiveness.

References

Ahmad, I. A. (2002), Handbook of applied econometrics and statistical inference, 1st edn, CRC Press, chapter Kernel Estimation in a Continuous Randomized Response Model.

Blair, G., Imai, K. & Zhou, Y.-Y. (2015), 'Design and analysis of the randomized response technique', *Journal of the American Statistical Association* **110**(511), 1304–1319.

Chambers, R. L. & Dunstan, R. (1986), 'Estimating distribution functions from survey data', *Biometrika* **73**(3), 597–604.

Diana, G. & Perri, P. (2011), 'A class of estimators for quantitative sensitive data', *Statistical Papers* **52**(3), 633–650.

Dubnicka, S. R. (2009), 'Kernel density estimation with missing data and auxiliary variables', *Australian & New Zealand Journal of Statistics* **51**(3), 247–270.

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

Folsom, R. E., Greenberg, B. G., Horvitz, D. G. & Abernathy, J. R. (1973), 'The two alternate questions randomized response model for human surveys', *Journal of the American Statistical Association* **68**(343), 525–530.

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R. & Horvitz, D. G. (1969), 'The unrelated question randomized response model: Theoretical framework', *Journal of the American Statistical Association* **64**(326), 520–539.

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

Gupta, S., Aloraini, B., Qureshi, M. N. & Khalil, S. (2020), 'Variance estimation using randomized response technique', *REVSTAT – Statistical Journal* **18**(2), 165 – 176.

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

Gupta, S., Kalucha, G. & Shabbir, J. (2017), 'A regression estimator for finite population mean of a sensitive variable using an optional randomized response model', *Communications in Statistics - Simulation and Computation* **46**(3), 2393–2405.

Gupta, S., Kalucha, G., Shabbir, J. & Dass, B. K. (2014), 'Estimation of finite population mean using optional rrt models in the presence of nonsensitive auxiliary information', *American Journal of Mathematical and Management Sciences* **33**(2), 147–159.

Gupta, S., Mehta, S., Shabbir, J. & Khalil, S. (2018), 'A unified measure of respondent privacy and model efficiency in quantitative rrt models', *Journal of Statistical Theory and Practice* **12**(3), 506–511.

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

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

Horvitz, D. G. & Thompson, D. J. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47**(260), 663–685.

Jones, E. E. & Sigall, H. (1971), 'The bogus pipeline: A new paradigm for measuring affect and attitude.', *Psychological Bulletin* **76**(5), 349–364.

Kalucha, G., Gupta, S. & Shabbir, J. (2016), 'A two-step approach to ratio and regression estimation of finite population mean using optional randomized response models', *Hacettepe Journal of Mathematics and Statistics* **45**, 1819 – 1830.

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

Khalil, S., Noor-ul Amin, M. & Hanif, M. (2018), 'Estimation of population mean for

a sensitive variable in the presence of measurement error', *Journal of Statistics and Management Systems* **21**(1), 81–91.

Khalil, S., Zhang, Q. & Gupta, S. (2021), 'Mean estimation of sensitive variables under measurement errors using optional rrt models', *Communications in Statistics - Simulation and Computation* **50**(5), 1417–1426.

MANGAT, N. S. & SINGH, R. (1990), 'An alternative randomized response procedure', *Biometrika* **77**(2), 439–442.

Mehta, S. & Aggarwal, P. (2018), 'Bayesian estimation of sensitivity level and population proportion of a sensitive characteristic in a binary optional unrelated question rrt model', *Communications in Statistics - Theory and Methods* **47**(16), 4021–4028.

Mostafa, S. A. & Ahmad, I. A. (2019), 'Kernel density estimation from complex surveys in the presence of complete auxiliary information', *Metrika* **82**(3), 295–338.

Nadaraya, E. (1964), 'On estimating regression', *Theory of Probability and Its Applications* **9**, 141–142.

Narjis, G. & Shabbir, J. (2020), 'Estimation of population proportion and sensitivity level using optional unrelated question randomized response techniques', *Communications in Statistics - Simulation and Computation* **49**(12), 3212–3226.

Parzen, E. (1962), 'On Estimation of a Probability Density Function and Mode', *The Annals of Mathematical Statistics* **33**(3), 1065 – 1076.

Pearson, K. (1894), 'Contributions to the mathematical theory of evolution', *Philosophical Transactions of the Royal Society of London. (A.)* **185**, 71–110. III.

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

Poole, W. K. (1974), 'Estimation of the distribution function of a continuous type random variable through randomized response', *Journal of the American Statistical Association* **69**(348), 1002–1005.

Raghavarao, D. & Federer, W. T. (1979), 'Block total response as an alternative to the randomized response method in surveys', *Journal of the Royal Statistical Society. Series B (Methodological)* **41**(1), 40–45.

Rao, J. N. K., Kovar, J. G. & Mantel, H. J. (1990), 'On estimating distribution functions and quantiles from survey data using auxiliary information', *Biometrika* **77**(2), 365–375.

Reynolds, W. (1982), 'Development of reliable and valid short forms of the marlowe-crowne social desirability scale', *Journal of Clinical Psychology* **38**, 119–125.

Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function', *The Annals of Mathematical Statistics* **27**, 832–837.

Scott, D. W. & Terrell, G. R. (1987), 'Biased and unbiased cross-validation in density estimation', *Journal of the American Statistical Association* **82**(400), 1131–1146.

Shou, W. & Gupta, S. (2023), 'Kernel density estimation using additive randomized response technique (rrt) models', *Communications in Statistics - Simulation and Computation* **0**(0), 1–10.

Sousa, R., Shabbir, J., Real, P. C. & Gupta, S. (2010), 'Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information', *Journal of Statistical Theory and Practice* **4**(3), 495–507.

Wand, M. P. & Jones, M. C. (1994), *Asymptotic MSE and MISE approximations*, CRC Press, p. 19–23.

Warner, S. L. (1965), 'Randomized response: A survey technique for eliminating evasive answer bias', *Journal of the American Statistical Association* **60**(309), 63–69.

Warner, S. L. (1971), 'The linear randomized response model', *Journal of the American Statistical Association* **66**(336), 884–888.

Watson, G. S. (1964), 'Smooth regression analysis', *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* **26**(4), 359–372.

Young, A., Gupta, S. & Parks, R. (2019), 'A binary unrelated-question rrt model accounting for untruthful responding', *Involve, a Journal of Mathematics* **12**, 1163–1173.