SEHRA, SUPRITI, M.A. Two-stage Optional Randomized Response Models. (2008) Directed by Dr. Sat Narain Gupta. 69pp.

Social desirability bias (SDB) is defined as a tendency in people to present themselves in a more socially acceptable light, when faced with sensitive questions. People with a higher degree of SDB tend to give answers that will make them look good rather than those that are accurate. Randomized Response Technique (RRT) is one of several techniques used by researchers to circumvent social desirability bias in personal interview surveys. Starting from the pioneering work of Warner (1965), many versions of RRT have been developed that can deal with both categorical and quantitative responses. In this thesis we will focus only on those RRT models that are useful for quantitative responses. We will discuss a variety of quantitative RRT models including full, partial and optional RRT models. However, our primary focus in this thesis will be on optional RRT models. Specifically we will compare one-stage optional RRT models with two-stage optional RRT models.

For optional RRT models, both additive and multiplicative RRT models have been used in the literature. However, survey respondents with minimal or no mathematical background may find additive models easier to handle. In this thesis we will discuss some other advantages of using additive optional RRT models as opposed to multiplicative optional RRT models. We will develop unbiased estimators for both the mean and the sensitivity level of a quantitative response sensitive question. We will also try to validate the proposed estimators by way of a simulation study. Throughout this thesis, we will use only the simple random sampling with replacement (SRSWR) design. However, the results can also be extended to other sampling designs.

TWO-STAGE OPTIONAL RANDOMIZED RESPONSE MODELS

by

Supriti Sehra

A Thesis 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
Master of Arts

Greensboro
2008

Approved by

Committee Chair

This thesis is dedicated to

My husband and our families
For all their support and encouragement.

APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair_____

Committee Members_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

I would like to thank my advisor and committee chair Professor Sat Gupta for all his guidance, support and patience throughout the preparation of this thesis. I would also like to thank thesis committee members Professor Maya Chhetri and Professor Scott Richter for their time and effort in reviewing this work.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER I

INTRODUCTION

1.1 Background

One of the classic problems encountered in behavioral and social sciences is estimating the prevalence of a sensitive or delicate behavior. For example, researchers may want to find out the proportion of people that have cheated on a test in high school, or may want to estimate the number of alcoholic drinks people consume per week. A method that is widely used in the estimation of this type of sensitive behavior often involves conducting some form of a survey. But before the investigators can assert the validity of responses in their survey, they need to overcome a major obstacle: a condition known as "Social Desirability Response Bias."

In brief, Social Desirability Bias (SDB) can be defined as the bias created due to a person's desire to be viewed favorably by others. Specifically, individuals have a natural inclination to present themselves in a more positive or *socially acceptable* manner to their interviewers, due to their perceived fear of negative consequences. These negative consequences may be concrete punitive actions or something intangible like embarrassment, depending on the type of sensitive question being asked. For example, if a sensitive question involves admitting to an illegal behavior, respondents may genuinely be afraid of repercussions from the law. By the same token, if the sensitive question entails answering personal health questions, respondents may be embarrassed to reveal their true behavior in fear of negative judgments by the interviewer. Regardless, researchers end up dealing with fabricated or distorted

responses, thus leading to inaccurate or invalid estimates of the sensitive behaviors they are studying. SDB usually presents itself in relation to external environmental situations or internal individual characteristics. External situations may include the effect of who is administering the survey or an item characteristic in terms of "perceived desirability of behavior." For example, a person may not want to admit to having called in sick to work without being sick. Internal individual attribute relates to differences between individual personality traits in terms of "impression management" or "self-deception" (Randall and Fernandes, 1991). For example, a person may be reluctant to admit that he has a drinking problem, perhaps even to himself, even though clinically he would be diagnosed as an alcoholic.

1.2 Circumventing Social Desirability Response Bias

There have been several techniques explored in the past to cope with SDB. We will briefly outline some of these techniques which include guaranteeing confidentiality, employing the SDB scale, and using the Bogus Pipe Line (BPL) method. Finally, we will describe in detail our chosen method of dealing with SDB: the Randomized Response Technique (RRT).

One of the best ways to elicit truthful responses, and thus minimize SDB, is to guarantee confidentiality. This may not be as simple as the researchers verbally assuring the respondents that "your data will only be analyzed in aggregate form, and you cannot be individually identified." To ensure that respondents will provide the most truthful responses possible, they have to *believe* that "complete anonymity" from everyone involved in the process actually exists. This is best achieved via a completely anonymous survey, like a mail-in questionnaire with no identifying data, where there is no respondent-to-researcher contact whatsoever. Unfortunately, mail-in surveys have abysmal response rates, thus leading to a dif-

ferent set of problems. Not only will the low response rate affect the precision of results, but the non-responders may fit into their own distinct categories, creating a non-response bias in its own right. On the flipside, in-person surveys have the highest response rates, undoubtedly because it is much more difficult for the respondents to refuse participation in a face-to-face situation with the interviewers. But this same face-to-face interaction also makes it more awkward for the respondents to be completely candid with their responses to sensitive questions. In-person studies having the most direct respondent-to-researcher contact, not surprisingly, tend to markedly accentuate SDB and produce the least reliable results.

The method just outlined tries to decrease SDB, where as other methods aim to measure and account for its severity, and then try to make adjustments accordingly during analysis. One such method uses a "SDB scale." Crowne and Marlowe (1960), developed the Crowne-Marlowe Social Desirability Bias (MCSDB) scale which measures individual SDB levels. A survey which lists 33 personal attitude type statements is administered to the subjects. The subjects are asked if each statement is true or false for them personally. Each statement has a socially accepted or morally "correct" answer, and depending on how the statement is worded, that could be "true" or "false (see example below). A socially correct answer is scored as one point, while an incorrect answer is scored as zero points. Thus the higher the SDB score for the whole survey, the greater the tendency of the individual to give socially desirable responses (or inflate SDB.) Although useful, the MCSDB scale measurement can add significantly to the respondent burden which can lead to lower response rates. To this end, Reynolds (1982) developed comparable 11, 12, and 13 item condensed versions of the test. The 13 item version yielded the best results in approximating the original MCSDB scale, and is shown in Table 1.1 below. The socially appropriate or correct answer is listed in parentheses next

to each item. These SDB scores can be used in conjunction with the other questions under investigation to see if there is a significant impact on how individuals answer other questions. If so, the SDB scores can be used as a covariate during the analysis phase of the study, to adjust the means for the sensitive question(s) under investigation.

Table 1.1: Reynolds 13-Point SDB Questionnaire with socially desirable responses shown in parentheses

1	It is sometimes hard for me to go on with my work if I am not encouraged. (F)
2	I sometimes feel resentful when I don't get my way. (F)
3	On a few occasions, I have given up doing something because I thought too little of my ability. (F)
4	There have been times when I felt like rebelling against people in authority even though I knew they were right. (F)
5	No matter who I'm talking to, I'm always a good listener. (T)
6	There have been occasions when I took advantage of someone. (F)
7	I'm always willing to admit it when I make a mistake. (T)
8	I sometimes try to get even, rather than forgive and forget. (F)
9	I am always courteous, even to people who are disagreeable. (T)
10	I have never been irked when people expressed ideas very different from my own. (T)
11	There have been times when I was quite jealous of the good fortune of others. (F)
12	I am sometimes irritated by people who ask favors of me. (F)
13	I have never deliberately said something that hurt someone's feelings. (T)

Another novel strategy for circumventing SDB is using an experimental method known as the Bogus Pipe Line (BPL) method. Developed by Jones and Sigall (1977), it may be best described as a "fake lie detector test". Respondents are hooked up to electrodes that supposedly send signals to a device which can detect if the respondent is being truthful or not. During the initial phase of the experiment, respondents are asked to answer some questions, the answers for which

are already known to the investigators. They are encouraged to give some truthful responses, and some that are fabricated, so that the investigators can "prove" to the respondents that the device to which they are attached actually works. Every time an untruthful response is provided, the interviewer presses a hidden buzzer indicating that an untruthful response has been detected. This technique can be quite successful in reducing SDB, since many if not most of the respondents do end up believing that the device is legitimate (Roese, and Jamieson, 1993). However, as with the SDB scale method described earlier, and even more so with this method, a major drawback is that it is extremely resource intensive, not only for the respondents, but also for the investigators. A great deal of time and money need to be invested in setting up and running the rather complex experimental conditions and physical machinery. Also, the method is obviously not portable. This method may be best used when the benefits of usage outweigh the large costs involved. Another pertinent philosophical criticism posed against this technique is that the researcher may not be acting in a truly ethical manner in conducting the study, as the researcher is essentially deceiving the subject as to the capabilities of the machine used.

One important note about the nature of SDB must be stressed. All the methods that we list in this thesis cannot completely circumvent SDB, they just try to mitigate it. As described earlier, people also have an internal "self-deception" mechanism continually at work, which tries improve their self-esteem. This results in people subconsciously putting themselves in a more positive light without being aware of it.

1.3 Randomized Response Technique

Now we will describe another method for coping with SDB, which will be the main focus of discussion in this thesis. This resourceful method was devised by Warner (1965) to decrease SDB and is called the Randomized Response Technique (RRT). Broadly speaking, while employing this technique, the respondent is asked to randomize or "scramble" the response to a sensitive or threatening question. This "scrambling" is based on some preset randomization device. The key point to take notice of here is that the subject's reported response is based on *chance*. That is to say, it is based on the outcome of the randomization device, and as a result, the respondent's "public" answer is "cloaked". The respondent is convinced that the unscrambling can be done only at the group level and not at the individual level. This allows the respondent to have the anonymity needed to be able to answer freely in a face-to-face situation, thereby helping decrease SDB.

A quick example of a typical RRT set-up will help clarify the technique better. Let's say we are interested in estimating the "average number of alcoholic drinks consumed daily" in a population. In a regular in-person survey, the interviewer would directly ask the respondent: "how many alcoholic drinks do you consume on a typical day?" This question would make some respondents feel quite uncomfortable, especially in a face-to-face situation with the interviewer, and therefore pressure them into understating or distorting their "true responses." But in the RRT realm, for the same question, the scenario would be somewhat different. Let's say we have chosen our randomization device to be a standard deck of cards, with the face cards discarded. The respondent would be given the deck of cards, asked to pick a card, then *without letting the interviewer see that card*, simply report the sum of the "number listed on the card" and the "true response" back to the interviewer.

It is important to reiterate that another reason RRT convinces respondents

give more truthful responses without embarrassment (albeit scrambled responses), is that they can easily see that the interviewers do not know which card they personally picked, and thus cannot know their true responses on an individual level. But an even more interesting aspect about using this method is that, even without knowing the true individual level responses, researchers can calculate group level estimates for the sensitive behavior being studied. This is done using the known distribution properties of the randomization device being employed and other mathematical techniques.

Most RRT models are some variant of the example described above. But all have two things in common: one is the "randomization" feature and the other is the "anonymity" feature – especially in front of the interviewer. This is where the choice of randomization devices becomes important. Randomization devices theoretically can take various forms, for example, a coin, a pair of dice, a deck of cards, a random number generator, or even a roulette wheel. But practically speaking, a randomization device that is portable, easy to understand and handle for respondents in any RRT experimental environment, would be preferable. More importantly, a randomization device that *easily* allows the respondent to hide the outcome of the device from the interviewer would be ideal. A standard or appropriately modified deck of cards fits all the above criteria. Thus in most of the models described in this thesis we will use a deck of cards as our preferred randomization device.

Many types of RRT models have been developed in the past, and they can be categorized as two major types: binary and quantitative. Binary response models are used to estimate the *proportion* of some behavior or occurrence in a population, and elicit a binary response during the RRT process. For example, to estimate the "proportion of people who drank coffee today" we could list statements: "I drank coffee today" and "I did not drink coffee today" on the cards in a deck.

The respondent would randomly pick a card, and simply respond "true" or "false" to the statement listed on the card. Thus the respondent is not stating explicitly whether or not he drank coffee, he is simply responding to the statement shown on the randomly drawn card. Quantitative response models are used to estimate the *mean* value of some behavior in a population. They can be further sub-classified as either additive models or multiplicative models. For example, to estimate "average number of cups of coffee consumed daily" we could use a standard deck of cards. The respondent would pick a card, and for the additive model, would respond with the sum of the card value and his true response; while for the multiplicative model, he would respond with the product of the card value and his true response.

RRT models can also be categorized by how the respondents are instructed to randomize. If all respondents are asked to randomize their response, the model is characterized as a "full randomization model." If some of respondents are instructed to randomize their response, the model is characterized as a "partial randomization model" or a "two-stage model." And finally, if respondents are given an option to randomize their response, it is characterized as an "optional randomization model." In connection with the "optional randomization model," another key concept of "question sensitivity level" needs introduction. "Sensitivity" is defined as the proportion of respondents who think the question is sensitive and hence choose to randomize their responses. Earlier we had specified that in quantitative RRT models, we estimate the mean value for a sensitive behavior. But in an optional randomization model, we end up estimating two parameters: the mean value of the sensitive behavior and the "sensitivity" level of that behavior. RRT models can also be classified as one-stage and two-stage (as indicated above, with the "partial randomization model". One-stage specifies respondents to follow either the full randomization or optional randomization route in one phase. On the other hand,

"partial randomization models" are like two-stage models, because a proportion of respondents (based on some randomization device) is asked to answer truthfully, while the rest are asked to randomize (again based on some randomization device.)

1.4 Goals

In general, each model described above in progression is better than the one previous to it, in terms of a balance between the model's relative efficiency (decreased variance) and model's ease of use in conjunction with the respondent's perceived anonymity level. Previously, a one-stage optional additive model using two independent samples (Gupta et al., 2006) and a two-stage optional multiplicative model using one sample (Gupta and Shabbir, 2007) have been specified. Both have their own distinct advantages and disadvantages. We aim to use these two models as a basis, and propose a "two-stage additive optional model using two independent samples" which would optimize the benefits of each, and produce a model with greater efficiency.

In Chapter 2, we will start with the theoretical framework of RRT and describe the methodology involved in various previous models that are relevant for the development of our current models. Namely, we will begin by outlining the basic binary and quantitative response models. In Chapter 3, we will introduce the optional RRT model. We will also compare and contrast the differences among full, partial, and optional quantitative randomization models. Moreover, we will discuss the shortcomings and benefits of additive versus multiplicative models in conjunction with one-stage and two-stage studies. In Chapter 4, we will specify in detail our proposed "two-stage optional additive RRT model using two independent samples." In Chapter 5, we will provide simulation results for this model in comparison to other models and also provide some concluding remarks.

We would like to note that there are many other types of RRT models that exist in the literature. However, our focus in this thesis will be on variations of Warner's (1965) binary response RRT model and the Eichhorn and Hayre (1983) quantitative response model.

CHAPTER II

ONE AND TWO-STAGE RANDOMIZED RESPONSE MODELS

In this chapter we will introduce both binary response and quantitative response RRT models. First we will outline some binary response models, since they form the historical backbone of RRT research. But after that we will concentrate mainly on the quantitative models, since they form the basis of the proposed models in this thesis.

As briefly introduced earlier, binary response models are used in estimating the proportion of some behavior or occurrence in the population. As the name implies, there are two mutually exclusive responses possible ('yes/no", "agree/disagree", "true/false" etc.) On the other hand, quantitative response models involve numeric responses. Thus, if we want to estimate the mean of a random variable that describes some behavior or occurrence, we would use a quantitative response model. In the following sections, we will describe in detail the early models developed for these two situations. In conjunction, we will also define and explain the differences between full and partial randomization models, which will come into play when later defining our proposed models.

2.1 Binary Randomized Response Models

2.1.1 Full Binary RRT Model

Most of the RRT models are based on the pioneering work done by Stanley Warner in 1965. Warner's model is used to estimate the proportion of subjects with a

sensitive behavior or characteristic. In order to estimate this proportion, a randomization device such as a deck of flash cards is employed, whereby all subjects are asked to randomize their responses based on the deck of cards. The sampling scheme considered here and throughout this thesis is simple random sampling with replacement. Since all subjects are asked to scramble their responses, the model is called a "full randomization" model. Two types of statements are written on the cards: a certain proportion of cards state "I have characteristic A;" the remaining proportion of the cards state "I do not have characteristic A." If the subject agrees with the statement on the card that he happens to pick, the response would be "yes." Conversely, if the subject disagrees with the statement, the response would be "no." It is important to note, that the interviewer has not seen the card that the respondent has picked, and thus does not know which question is being answered. The interviewer simply records the "yes" or "no" responses.

So mathematically, letting π be the true proportion of the subjects with the sensitive characteristic, and p be the proportion of cards with "I have characteristic A" written on them, the probability of a "yes" response, p_y, would be

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

Solving for π , we get

$$\pi = \frac{p_y - (1-p)}{2p-1}, \qquad p \neq .5 \tag{2.2}$$

Using the sample data leads to Warner's unbiased estimator for π given by

$$\widehat{\pi}_w = \frac{\widehat{p}_y - (1-p)}{2p-1}. \tag{2.3}$$

Note the proportion of "yes" responses is estimated by

$$\widehat{p}_y = \frac{n_1}{n}, \tag{2.4}$$

where n is the sample size and n_1 is the number of "yes" responses.

From the fact that

$$Var(\widehat{p}_y) = \frac{p_y\,(1-p_y)}{n}, \tag{2.5}$$

it can be verified that the variance for the estimator $\widehat{\pi}_w$ is:

$$Var\,(\widehat{\pi}_w) = \frac{\pi\,(1-\pi)}{n} + \frac{p\,(1-p)}{n\,(2p-1)^2}. \tag{2.6}$$

The second term in the above equation $\left(\frac{p(1-p)}{n(2p-1)^2}\right)$ is the penalty for using the RRT model. Note that this penalty is minimized when $p \approx 0$ or $p \approx 1$.

For example, if we were interested in estimating the proportion of people who have ever tried cocaine, the following display shows the four possible outcomes for this RRT process. The respondent's reported response, which is listed in the cells, is based on two factors: his "true cocaine user status" and the "card the respondent he happens to pick". The "true cocaine user status" is specified in the left margin column, while the "card picked" is specified in the header row. The proportions for each of the two types of cards and the proportions for each of the two types of user status are shown in parentheses

True Status	Card Picked	
	I have tried cocaine(p)	I have never tried cocaine($1-p$)
User(π)	yes	no
Non-User($1-\pi$)	no	yes

Each individual participant would simply answer "yes" or "no" truthfully, based on the card he happens to pick. So if a person who has tried cocaine before picks the "I have tried cocaine" card, he would agree with that statement, so he would simply answer "yes." On the other hand, if a person who has tried cocaine

picks the "I have never tried cocaine" card, he would disagree with that statement, so he would answer "no." Using the display, it's easy to verify the specification for proportion of "yes" responses listed in equation (2.1) above. Note that the proportions p and $1-p$ are known, as are the number of "yes" responses n_1 and the sample size n, hence we can calculate the values of $\widehat{\pi}_w$ and $Var\left(\widehat{\pi}_w\right)$.

2.1.2 Partial Binary RRT Model

Building on Warner's original model, Mangat and Singh (1990) proposed a slightly different model to increase the efficiency of Warner's estimator. Rather than having all subjects scramble their response, they proposed a two-stage model. In stage 1, some proportion of the subjects would respond truthfully, while the rest would go to stage 2, whereby they scramble their response (i.e., stage 2 subjects would follow Warner's model.) Which subjects go to stage 2 is based on a "stage 1 randomization device". For instance, the "stage 1 randomization device" could be a deck of cards, with a known proportion (T) of cards asking the subject to answer the "I have characteristic A" question truthfully; the remaining proportion $(1-T)$ of cards ask the subject to use the "stage 2 randomization device" to answer the question. All subjects relegated to stage 2 scramble their responses, exactly as they would have with Warner's full randomization model. Again, the interviewer does not know which question you answered or if you answered in stage 1 or stage 2; he merely records the "yes" or "no" responses.

Updating Warner's model above, let π be the true proportion of subjects with the sensitive characteristic, T be the proportion of subjects answering truthfully in stage 1, and p be the proportion of cards in stage 2 with "I have characteristic A" written on them, then the probability of a "yes" response, p_y, becomes

$$p_y = T\pi + (1-T)\left\{p\pi + (1-p)(1-\pi)\right\}. \tag{2.7}$$

Using the sample data leads to the Mangat and Singh unbiased estimator for the proportion of sensitive characteristic, given by

$$\widehat{\pi}_{ms} = \frac{\widehat{p}_y - (1-T)(1-p)}{(2p-1) + 2T(1-p)}, \tag{2.8}$$

where $\widehat{p}_y = \frac{n_1}{n}$, n is the sample size, and n_1 is the number of "yes" responses.

The variance for this estimator is given by

$$Var(\widehat{\pi}_{ms}) = \frac{\pi(1-\pi)}{n} + \frac{(1-T)(1-p)[1-(1-T)(1-p)]}{n[(2p-1)+2T(1-p)]^2}. \tag{2.9}$$

The second term in the equation is the penalty for using this model. Note that when $T = 0$ this model reduces to Warner's full RRT model specified above.

So if we use the same example of estimating the proportion of people who have ever tried cocaine, the following display shows all the possible reported responses using the partial RRT model:

True Status	Stage 1 Card (T)	Stage 2 Card ($1-T$)	Reported Response
User (π)	Tell the truth	N/A	YES
	go to S2	I've tried cocaine	YES
	go to S2	I've never tried cocaine	NO
Non-User ($1-\pi$)	Tell the truth	N/A	NO
	go to S2	I've tried cocaine	NO
	go to S2	I've never tried cocaine	YES

Note that the Mangat and Singh partial RRT estimator $\widehat{\pi}_{ms}$ is more efficient than Warner's full RRT estimator $\widehat{\pi}_w$ for fixed sample size if

$$T > \frac{1-2p}{1-p},$$

which is always true when $p > .5$.

2.2 Quantitative Additive Randomized Response Models

While the previously described binary response RRT models form a good basis for explaining the intricacies of how RRT models function, we will devote the rest of the thesis to quantitative response RRT models. For quantitative response RRT models we want to estimate the mean prevalence of some sensitive behavior. We will outline the additive models first, since they are easier to comprehend for the respondents and also, mathematically easier to analyze. For the additive quantitative model, subjects would be asked to scramble their responses using a randomization device like a deck of cards. Now, each of the cards in the deck would have a number listed on it, where the numbers in the deck follow a known probability distribution. The subject would be asked to add his "true response" to the "number listed on card" he picked, and then report only the *sum* to the interviewer. The interviewer cannot see the card picked and would simply record a number.

2.2.1 Full Additive RRT Model

The quantitative additive version of Warner's binary full randomization model is specified below (Warner, 1971). Again, since it's a full randomization model, all subjects are instructed to randomize their responses based on a outcome of the randomization device with known mean and variance.

Mathematically, for the additive model, if we let Y be the reported response, X be the true sensitive variable of interest with *unknown* mean μ_x and *unknown* variance σ_x^2, and S be the scrambling variable (independent of X) with *known* true mean μ_s ($E(S) = \mu_s$) and *known* variance σ_s^2, then

$$Y = X + S. \tag{2.10}$$

The expected response is given by

$$\begin{aligned} E(Y) &= E(X) + E(S) \\ &= \mu_x + \mu_s. \end{aligned} \tag{2.11}$$

Estimating $E(Y)$ by the sample mean $\overline{Y}$ of the reported responses, we get the unbiased estimator for the mean of the sensitive variable, given by

$$\widehat{\mu}_x = \overline{Y} - \mu_{s.} \tag{2.12}$$

The variance for this estimator is given by

$$\begin{aligned} Var\left(\widehat{\mu}_x\right) &= Var\left(\overline{Y}\right) = \frac{\sigma_y^2}{n} \\ &= \frac{\sigma_x^2}{n} + \frac{\sigma_s^2}{n}. \end{aligned} \tag{2.13}$$

In equation (2.13), $\frac{\sigma_s^2}{n}$ represents the penalty for using a RRT model.

For example, if researching the average weekly alcohol consumption rate, we may ask the subjects "how many alcoholic drinks do you typically consume per week?" We would ask them to add their true response to the numeric card they pick from the deck. The cards in the deck we created list the numbers $5, 6, 7, 8,$ and 9 with corresponding frequencies of $6, 10, 16, 10,$ and 6, respectively. Thus, the frequency distribution of the card values have a mean $\mu_s = 7$ and variance of 1.42. If we get the following experimental result corresponding to a sample of size 5,

TRUE # of drinks	Card Picked	Reported Response
5	9	14
0	7	7
7	5	12
6	6	12
3	6	9
		$\overline{Y}$= 10.8
		$\widehat{\mu}_x$=$\overline{Y}$−μ_s= $10.8 - 7 = 3.8$

then our estimate for μ_x would be 3.8.

2.2.2 Partial Additive RRT Model

Building upon the quantitative full additive RRT model, Gupta and Thornton (2004) have described a partial (two-stage) quantitative randomization model. Similar to the binary version of the partial randomization model, some known proportion (T) of cards in the deck ask subjects to respond truthfully (stage 1), while the remaining proportion $(1-T)$ of cards ask the respondents to report the sum of the number listed on the card and their true response (stage 2).

Let T be the proportion of cards asking respondents to answer truthfully, Y be the reported response, be the sensitive variable of interest with *unknown* mean μ_x and *unknown* variance σ_x^2, and S be the scrambling variable (independent of X) with *known* true mean μ_s $(E(S) = \mu_s)$ and variance σ_s^2, then

$$Y = \begin{cases} X & \text{with probability } T \\ X+S & \text{with probability } (1-T). \end{cases} \tag{2.14}$$

Expected response is given by

$$\begin{aligned} E(Y) &= (T)E(X) + (1-T)E(X+S) \\ &= (T)\mu_x + (1-T)(\mu_x + \mu_s) \\ &= \mu_x + (1-T)\mu_s. \end{aligned} \tag{2.15}$$

Substituting the sample mean $\overline{Y}$ of the reported responses for $E(Y)$, we get the unbiased estimator for the mean of the sensitive variable given by

$$\widehat{\mu}_x = \overline{Y} - (1-T)\mu_s. \tag{2.16}$$

The variance for this estimator is given by

$$Var\left(\widehat{\mu}_x\right) = \frac{\sigma_x^2}{n} + \frac{(1-T)(\sigma_s^2 + T\mu_s^2)}{n}$$

$$= \frac{1}{n}[\sigma_x^2 + (1-T)(\sigma_s^2 + T\mu_s^2)]. \quad (2.17)$$

Using our previous example of "how many alcoholic drinks do you typically consume per week?", we now change it to a two-stage model, where 20% ($T = .2$) are asked to answer truthfully (stage 1). The rest, 80% ($1 - T = .8$) are asked to randomize their response (stage 2). The cards in the second part of the deck list the numbers $5, 6, 7, 8$,and 9 with corresponding frequencies of $6, 10, 16, 10$, and 6, respectively Thus, the frequency distribution of the card values have a mean $\mu_s = 7$ and variance of 1.42. If we get following experimental result corresponding to a sample of size 5, then our estimate for μ_x would be 4.

	$T = .2$	$(1-T) = .8$	
True # of drinks	Stage 1	Stage 2	Reported Response
5	go to S2	9	14
0	go to S2	7	7
7	go to S2	5	12
6	Tell Truth	N/A	6
3	go to S2	6	9
			$\overline{Y} = 9.6$
		$\widehat{\mu}_x = \overline{Y} - (1-T)\mu_s = 9.6 - (.8)7 = 4$	

Gupta and Thornton (2002) noted that for a quantitative response, the variance using the partial randomization model is less than the variance of the full randomization model if

$$T > \frac{\mu_s^2 - \sigma_s^2}{\mu_s^2} = 1 - {CV_s}^2,$$

where CV_s is the coefficient of variation of S.

2.3 Quantitative Multiplicative Randomized Response Models

Similar to the additive quantitative RRT models, multiplicative models also estimate the mean prevalence of a sensitive behavior. Again, a deck of cards with known probability distribution is employed, but now when the subjects scramble their responses, they are asked to report the *product* of the "true response" and the "number listed on the card" picked. Of course, the interviewer simply records a number and cannot see the card picked.

2.3.1 Full Multiplicative RRT Model

The full randomization model was developed by Eichhorn and Hayre (1983). It is analogous to the full randomization additive model.

For the multiplicative model, let Y be the reported response, X be the true sensitive variable of interest with *unknown* mean μ_x and *unknown* variance σ_x^2, and S be the scrambling variable (independent of X) with *known* true mean $E(S) = \mu_s = \theta$ and *known* variance σ_s^2, the respondent is asked to report

$$Y = \frac{XS}{\theta}. \tag{2.18}$$

Expected response is given by

$$E(Y) = \frac{E(X)E(S)}{\theta} = \mu_x. \tag{2.19}$$

Estimating μ_x using the sample mean of the reported responses $\overline{Y}$, we get the unbiased estimator for the mean of the sensitive variable, given by

$$\widehat{\mu}_x = \overline{Y}. \tag{2.20}$$

The variance of this estimator is given by

$$Var\left(\widehat{\mu}_x\right) = \frac{1}{n}\left[\sigma_x^2 + \frac{\sigma_s^2}{\theta^2}\left(\sigma_x^2 + \mu_x^2\right)\right]. \tag{2.21}$$

2.3.2 Partial Multiplicative RRT Model

Similar to the additive partial randomization model, some known proportion (T) of cards in the deck ask subjects to respond truthfully (stage 1), while the remaining proportion ($1-T$) of cards ask the respondents to report the product of the number listed on the card and their true response (stage 2) divided by θ.

Mathematically, for the multiplicative model, let Y be the reported response, X be the sensitive variable of interest with *unknown* mean μ_x and *unknown* variance σ_x^2, and S be the scrambling variable (independent of X) with *known* true mean $E(S) = \mu_s = \theta$ and variance σ_s^2, then

$$Y = \begin{cases} X & \text{with probability } T \\ \frac{XS}{\theta} & \text{with probability } (1-T). \end{cases} \tag{2.22}$$

Expected response is given by

$$\begin{aligned} E(Y) &= (T)E(X) + (1-T)\frac{E(X)E(S)}{\theta} \\ &= (T)\mu_x + (1-T)\mu_x \\ &= \mu_x. \end{aligned} \tag{2.23}$$

Estimating μ_x using the sample mean of the reported responses $\overline{Y}$, we get the unbiased estimator for the mean of the sensitive variable, given by

$$\widehat{\mu}_x = \overline{Y}. \tag{2.24}$$

To calculate the variance for this estimator $\widehat{\mu}_x$, we must first calculate $Var\,(Y)$ since

$$Var\,(\widehat{\mu}_x) = Var\left(\overline{Y}\right) = Var\left(\frac{Y}{n}\right). \tag{2.25}$$

Noting that

$$E(X^2) = \sigma_x^2 + \mu_x^2 \tag{2.26}$$

and

$$E(S^2) = \sigma_s^2 + \theta^2, \tag{2.27}$$

we get

$$\begin{aligned} Var\,(Y) &= E(Y^2) - [E(Y)]^2 \\ &= \left[(T)E(X^2) + (1-T)\frac{E(X^2)E(S^2)}{\theta^2}\right] - \mu_x^2 \\ &= (T)(\sigma_x^2 + \mu_x^2) + (1-T)\frac{(\sigma_x^2 + \mu_x^2)(\sigma_s^2 + \theta^2)}{\theta^2} - \mu_x^2 \\ &= \sigma_x^2 + (1-T)\frac{\sigma_s^2}{\theta^2}\left(\sigma_x^2 + \mu_x^2\right). \end{aligned} \tag{2.28}$$

Finally, via equation (2.25) we can calculate $Var\,(\widehat{\mu}_x)$:

$$Var\,(\widehat{\mu}_x) = \frac{1}{n}\left[\sigma_x^2 + (1-T)\frac{\sigma_s^2}{\theta^2}\left(\sigma_x^2 + \mu_x^2\right)\right]. \tag{2.29}$$

If we compare the variance of the partial multiplicative and full multiplicative RRT models, the partial RRT model variance given by equation (2.29) will be lower than full RRT model variance given by equation (2.21),

$$\begin{aligned} \text{if} \quad \frac{1}{n}\left[\sigma_x^2 + (1-T)\frac{\sigma_s^2}{\theta^2}\left(\sigma_x^2 + \mu_x^2\right)\right] &< \frac{1}{n}\left[\sigma_x^2 + \frac{\sigma_s^2}{\theta^2}\left(\sigma_x^2 + \mu_x^2\right)\right] \\ \text{or if} \quad 1 - T &< 1. \end{aligned}$$

It should be noted that $1 - T < 1$ is always true, since $0 \leq T \leq 1$.

CHAPTER III

OPTIONAL RANDOMIZED RESPONSE MODELS

The key difference between the RRT models outlined thus far and the Optional RRT models we will discuss next is that now, the *respondent gets to decide* if he wants to give the true response or a scrambled response. The respondent is asked to answer truthfully if he considers the question non-sensitive, otherwise the respondent is instructed to provide a scrambled response. In earlier models, the subject was not given a choice on how to respond – either he was asked to give a true response or he was asked to give a scrambled response (although that was still dependent on chance). This choice gives the respondents an even higher sense of security with respect to the anonymity issue, thus encouraging more truthful responses.

3.1 One-stage Optional Randomized Response Models

One-stage optional RRT models are characterized by the fact that all subjects are allowed to choose whether they want to scramble their response or answer truthfully.

3.1.1 One-stage Multiplicative Optional RRT Model using One Sample

Gupta et al. (2002) developed a model in which the respondents are given the following two options:

a) Report a truthful response if you do not consider the question to be sensitive

b) Report a multiplicatively scrambled response if you consider the question to be sensitive

Based on this model, they called the proportion of subjects that scramble the response the *"sensitivity level"* of the behavioral question being studied. Note, now we have two parameters that need estimation – the usual sensitive question mean (μ_x) and the newly added sensitivity level (W).

Thus, let W $(0 \leq W \leq 1)$ be the sensitivity level (i.e., the proportion of respondents in the population who consider the question to be sensitive) and Y be a random variable where

$$Y = \begin{cases} 1 & \text{if response is scrambled} \\ 0 & \text{if response not scrambled.} \end{cases}$$

Note that $Y \sim Bernoulli(W)$ and

$$E(Y) = W. \tag{3.1}$$

Now the mathematical optional RRT model can be fully specified. Let X be the true response variable, and S be the scrambling variable (independent of X) with *known* true mean $E(S) = \mu_s = 1$, then the reported response Z is given by

$$Z = S^Y X. \tag{3.2}$$

Thus

$$\begin{aligned} E(Z) &= E(S^Y X) \\ &= E\left[S^Y X | Y = 1\right] P(Y = 1) + E\left[S^Y X | Y = 0\right] P(Y = 0) \\ &= E(S)E(X)P(Y = 1) + E\left[X\right] P(Y = 0) \\ &= \mu_x W + \mu_x(1 - W), \qquad \text{since } E\left[S\right] = 1 \\ &= \mu_x. \end{aligned} \tag{3.3}$$

Estimating $E(Z)$ using the sample mean of the reported responses $\frac{1}{n}\sum_{i=1}^{n} Z_i = \overline{Z}$, we get the unbiased estimator for the mean of the sensitive variable given by

$$\widehat{\mu}_x = \overline{Z}. \tag{3.4}$$

The variance for this estimator is given by

$$Var\left(\widehat{\mu}_x\right) = Var\left(\overline{Z}\right) = \frac{1}{n}\left[\sigma_x^2 + W\sigma_s^2\left(\sigma_x^2 + \mu_x^2\right)\right]. \tag{3.5}$$

Note that $Var\left(\widehat{\mu}_x\right)$ increases as W increases from 0 to 1. This makes sense since W represents the sensitivity of the question, and the greater the proportion of respondents that scramble their response, the smaller the number that answer the sensitive question truthfully, thus decreasing the estimation efficiency. Finally, in comparison to Eichhorn and Hayre's (1983) full multiplicative RRT model, where the variance equation (2.21) is given by

$$Var\left(\widehat{\mu}\right) = \frac{1}{n}\left[\sigma_x^2 + \frac{\sigma_s^2}{\theta^2}\left(\sigma_x^2 + \mu_x^2\right)\right],$$

the relative efficiency of the optional RRT model (if $\theta = 1$) is given by:

$$RE = \frac{\sigma_x^2 + \sigma_s^2\left(\sigma_x^2 + \mu_x^2\right)}{\sigma_x^2 + W\sigma_s^2\left(\sigma_x^2 + \mu_x^2\right)}. \tag{3.6}$$

Note that $RE \geq 1$ since $0 \leq W \leq 1$.

The second parameter (W) can be estimated as follows. Starting with the optional RRT model $Z = S^Y X$, we get

$$\begin{aligned}
\ln(Z) &= Y\ln(S) + \ln(X) \\
E\{\ln(Z)\} &= E(Y)E\{\ln(S)\} + E\{\ln(X)\} \\
E\{\ln(Z)\} &\approx W \cdot E\{\ln(S)\} + \ln\{E(X)\} \\
W &\approx \frac{E\{\ln(Z)\} - \ln\{\mu_x\}}{E\{\ln(S)\}}
\end{aligned} \tag{3.7}$$

This leads to the following estimator

$$\widehat{W} = \frac{\frac{1}{n}\sum_{i=1}^{n}\ln(Z_i) - \ln\left\{\frac{1}{n}\sum_{i=1}^{n} Z_i\right\}}{\delta}. \tag{3.8}$$

where $\delta = E\{\ln(S)\}$ is the known expected value of the log of the scrambling variable. Note that the above estimate is based on the first order Taylor's approximation of $E[\ln(X)]$ in the sense that $E[\ln(X)]$is approximated by $\ln[E(X)]$. Gupta and Shabbir (2007) use a second order approximation in the two-stage optional RRT model (discussed later.)

3.1.2 One-stage Multiplicative Optional RRT Model using Two Independent Samples

One possible drawback of the Gupta et al. (2002) model is that it cannot work with qualitative response variables. Thus Gupta and Shabbir (2004) developed a model to overcome this issue. Using a design similar to the Greenberg et al. (1969) unrelated question model they have used two independent samples.

Let X be the sensitive variable with mean μ_x and variance σ_x^2, and W be the sensitivity level. Furthermore, we use two independent random samples of sizes n_i $(i = 1, 2)$, where sample members in the ith sample use randomization device R_i. The two randomization devices follow *different* probability distributions with different means θ_i and different variances σ_{Si}^2. As with the one sample model, the respondent is given a choice to scramble or not. Namely, the respondent is asked to provide the truthful response if the question posed is not considered sensitive, otherwise the respondent is asked to *multiply* the true response with the output from the randomization device and report the final scrambled response. The reported

response Z_i $(i = 1, 2)$ for the ith sample, would be

$$Z_i = \begin{cases} X & \text{with probability } (1 - W) \\ S_i X & \text{with probability } W \end{cases}$$

where X is the true response, S_i $(i = 1, 2)$ is the scrambling variable corresponding to scrambling device R_i. We assume X, S_1, and S_2 to be mutually independent.

Noting that $E(S_i) = \theta_i$, we get

$$\begin{aligned} E(Z_i) &= E(X)(1 - W) + E(S_i X)W \\ &= \mu_x[1 + W(\theta_i - 1)], \qquad i = 1, 2. \end{aligned}$$

Then solving these two equations simultaneously leads to the following estimators:

$$\widehat{\mu}_x = \frac{\overline{Z}_1(\theta_2 - 1) - \overline{Z}_2(\theta_1 - 1)}{\theta_2 - \theta_1}, \qquad \theta_1 \neq \theta_2 \tag{3.9}$$

and

$$\widehat{W} = \frac{\overline{Z}_2 - \overline{Z}_1}{\overline{Z}_1(\theta_2 - 1) - \overline{Z}_2(1 - \theta_1)}, \tag{3.10}$$

where $E(Z_i)$ is estimated by the usual sample mean of the reported responses of the ith sample $(\overline{Z}_i)$.

Gupta and Shabbir (2004) derived the variances for the above estimators. The variance of $\widehat{\mu}_x$ is

$$Var\left(\widehat{\mu}_x\right) = \frac{1}{(\theta_2 - \theta_1)^2}\left[(\theta_2 - 1)^2\left(\frac{\sigma^2_{Z_1}}{n_1}\right) + (\theta_1 - 1)^2\left(\frac{\sigma^2_{Z_2}}{n_2}\right)\right], \tag{3.11}$$

where

$$\sigma^2_{Zi} = (\sigma^2_x + \mu^2_x)[1 - W + W(\theta^2_i + \sigma^2_{Si})] - \mu^2_x[\theta_i W + (1 - W)]^2, \tag{3.12}$$

and the variance for estimator $\widehat{W}$ is given by

$$Var(\widehat{W}) \approx \frac{1}{(\theta_2 - \theta_1)^2 \mu^2_x}\left[[1 + W(\theta_2 - 1)]^2\left(\frac{\sigma^2_{Z_1}}{n_1}\right) + [1 + W(\theta_1 - 1)]^2\left(\frac{\sigma^2_{Z_2}}{n_2}\right)\right]. \tag{3.13}$$

3.1.3 One-stage Additive Optional RRT Model using Two Independent Samples

To help resolve the problem of approximation used in calculating the $Var(\widehat{W})$ above, Gupta et al. (2006) developed another optional RRT model using two independent samples, but this time using an additive scrambling model rather than a multiplicative one.

As with the previous model, let X be the sensitive variable with mean μ_x and variance σ_x^2 and W be the sensitivity level. We select two independent random samples, of sizes n_i $(i = 1, 2)$, where the ith sample uses randomization device R_i, with mean θ_i and variance σ_{Si}^2. Again, the respondent is given a choice of scrambling the response if the question is considered sensitive. This time, rather than the product, the respondent who scrambles the response is asked to *add* the true response with the output from the randomization device and report the final scrambled response.

The reported response Z_i $(i = 1, 2)$ for the ith sample, would be

$$Z_i = \begin{cases} X & \text{with probability } (1 - W) \\ X + S_i & \text{with probability } W \end{cases}$$

where X is the true response, S_i $(i = 1, 2)$ is the variable value picked using scrambling device R_i and W is the sensitivity. We assume X, S_1, and S_2 to be mutually independent.

Note that

$$E(Z_i) = \mu_x + \theta_i(W) \qquad\qquad i = 1, 2.$$

Solving these two equations simultaneously leads to the estimators

$$\widehat{\mu}_x = \frac{\overline{Z}_1\theta_2 - \overline{Z}_2\theta_1}{\theta_2 - \theta_1}, \qquad \theta_1 \neq \theta_2, \tag{3.14}$$

and

$$\widehat{W} = \frac{\overline{Z}_2 - \overline{Z}_1}{\theta_2 - \theta_1}. \tag{3.15}$$

It can be verified that $\widehat{\mu}_x$ and $\widehat{W}$ are unbiased estimators of the true population mean μ_x and the true population sensitivity W, respectively.

Variances of these estimators are given by

$$Var\left(\widehat{\mu}_x\right) = \frac{1}{(\theta_2 - \theta_1)^2}\left[\theta_2^2\left(\frac{\sigma_{Z_1}^2}{n_1}\right) + \theta_1^2\left(\frac{\sigma_{Z_2}^2}{n_2}\right)\right] \tag{3.16}$$

and

$$Var(\widehat{W}) = \frac{1}{(\theta_2 - \theta_1)^2}\left[\frac{\sigma_{Z_1}^2}{n_1} + \frac{\sigma_{Z_2}^2}{n_2}\right]. \tag{3.17}$$

As compared to the multiplicative model, notice that the $Var\left(\widehat{W}\right)$ is not based on an approximation, since $\widehat{W}$ in equation (3.15) is no longer a ratio of two random variables, as was the case in equation (3.10).

Optimum sample sizes were also calculated that would *minimize* the sum of the estimator variances $\left[Var\left(\widehat{\mu}_x\right) + Var(\widehat{W})\right]$. These optimal values are given by

$$n_1 = \frac{n\sigma_{Z_1}\sqrt{\theta_2^2 + 1}}{\sigma_{Z_1}\sqrt{\theta_2^2 + 1} + \sigma_{Z_2}\sqrt{\theta_1^2 + 1}} \tag{3.18}$$

and

$$n_2 = \frac{n\sigma_{Z_2}\sqrt{\theta_1^2 + 1}}{\sigma_{Z_1}\sqrt{\theta_2^2 + 1} + \sigma_{Z_2}\sqrt{\theta_1^2 + 1}}. \tag{3.19}$$

In addition to being able to provide an exact expression for $Var(\widehat{W})$, this model affords the respondents a much more user friendly calculation (Singhal, 2004).

3.2 Two-stage Optional Randomized Response Models

As described earlier, a partial RRT model, where a known proportion of the respondents are asked to answer truthfully, generally leads to a better estimate for the mean of the sensitive question. We also learned that optional models help in respondent confidence of anonymity of the survey since they leave the scrambling at the discretion of the respondent. Using these two facts lead to the development of a "two-stage optional model." A two-stage optional RRT model is a combination of partial RRT model and optional RRT model. In "stage 1" a certain proportion of people are asked to respond truthfully; the rest go to "stage 2" where they follow the one-stage optional RRT model outlined above (i.e., they may scramble if they find the question to be sensitive enough.)

3.2.1 Attempted Two-stage Multiplicative Optional RRT Model using One Sample

To this end Ryu et al. (2006) outlined a model with the aim of producing a two-stage optional multiplicative RRT model. They have stipulated that in Stage 1, a randomly selected proportion of respondents (T) reply truthfully. The remaining respondents go to stage 2. In stage 2, a proportion P of these respondents *again* reply truthfully, the remaining respondents report the multiplicatively scrambled response: SX. In this model P *is assumed to be known*, thus only μ_x is being estimated. Nevertheless, Ryu et al. (2006) assume that their proportion, $(1-P)$, is the same as the W referenced in the Gupta et al. (2002) one-stage multiplicative RRT model.

Specifically, let Z be the reported response, X be the sensitive variable of interest with *unknown* mean μ_x and *unknown* variance σ_x^2, and S be the scrambling variable (independent of X). They show that the expected value of the reported

response Z is given by

$$E(Z) = E(X),$$

and an unbiased estimator of μ_x is given by

$$\widehat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} Z_i.$$

The variance of this estimator is given by

$$Var(\widehat{\mu}_x) = \frac{1}{n}\left[\sigma_x^2 + (1-T)(1-P)\,\sigma_s^2\left(\sigma_x^2 + \mu_x^2\right)\right].$$

Recalling that $(1-P)$ in their model is the same as W in the Gupta et al. (2002) model, they claim superiority of their estimator, since their $Var(\widehat{\mu}_x)$ is clearly less than or equal to the variance in the one-stage model given by

$$Var(\widehat{\mu}_x) = \frac{1}{n}\left[\sigma_x^2 + W\sigma_s^2\left(\sigma_x^2 + \mu_x^2\right)\right].$$

The problem is that this model assumes that W (or $(1-P)$ in this model) is a known quantity. This would not be a true two-stage *optional* RRT model, since only one unknown μ_x is being estimated (sensitivity level W or $(1-P)$ is not being estimated), and the whole point of an optional model is that the respondent makes a decision as to the sensitivity of the question posed. If he deems the question to be non-sensitive, he will answer truthfully in stage 2; if he deems the question to be sensitive, he will scramble the response. Predetermining who will tell the truth in stage 2 does not give this choice to the respondent.

3.2.2 True Two-stage Multiplicative Optional RRT Model using one sample

In response to the Ryu et al. (2006) model, Gupta and Shabbir (2007) developed a true "two-stage multiplicative optional RRT model with one sample" in which

two parameters are being estimated: the mean of the sensitive question μ_x and the sensitivity level W (i.e., the proportion of respondents who choose to scramble their reported response when given a choice.) For this model, some known proportion (T) of cards in the deck ask subjects to respond truthfully (stage 1), while the remaining cards ask respondents to follow the optional multiplicative RRT model. That is, the respondent is asked to provide the truthful response if he considers the question posed as non-sensitive. If he considers the question to be sensitive, he is asked to report the product of the number listed on the card and his true response.

Mathematically, for the two-stage multiplicative optional RRT model, we let Z be the reported response, X be the sensitive variable of interest with *unknown* mean μ_x and *unknown* variance σ_x^2, and S be the scrambling variable (independent of X) with *known* true mean $E(S) = \mu_s = 1$ and variance σ_s^2, then the reported response Z would be

$$Z = \begin{cases} X & \text{with probability } T + (1-T)(1-W) \\ SX & \text{with probability } (1-T)W. \end{cases} \tag{3.20}$$

If $V \sim Bernoulli(T)$, and $U \sim Bernoulli(W)$, then reported response can be written as

$$Z = \left\{X^V\right\}\left\{XS^U\right\}^{1-V}. \tag{3.21}$$

Taking expected values on both sides

$$\begin{aligned} E(Z) &= E(X).P(V=1) + E(X.S^U).P(V=0) \\ &= E(X).P(V=1) + E(X).\left\{E(S)P(U=1) + P(U=0)\right\}P(V=0) \\ &= E(X)P(V=1) + E(X)P(V=0) \qquad \text{since } E(S)=1 \\ &= E(X). \end{aligned}$$

Hence the mean of the sensitive variable can be estimated by

$$\widehat{\mu}_x = \frac{\sum_{i=1}^{n} Z_i}{n}. \tag{3.22}$$

The variance of the estimator $\widehat{\mu}_x$ is given by

$$Var\left(\widehat{\mu}_x\right) = Var(\overline{Z}) = \frac{1}{n}\left[\sigma_x^2 + (1-T)\,(W)\sigma_s^2\left(\sigma_x^2 + \mu_x^2\right)\right]. \tag{3.23}$$

The sensitivity level is estimated in a similar manner as in the Gupta et al. (2002) one-stage multiplicative model using a single sample, except that a second order Taylor's approximation is used here rather than first order. Taking the natural log of both sides in equation (3.21), we get

$$\begin{aligned} \ln(Z) &= V.\ln(X) + (1-V)\{\ln(X) + U\ln(S)\}. \\ &= \ln(X) + (1-V).U.\ln(S). \end{aligned} \tag{3.24}$$

Taking the expected values of both sides, we get

$$\begin{aligned} E[\ln(Z)] &= E[\ln(X)] + E(1-V).E(U).E[\ln(S)] \\ &= E[\ln(X)] + (1-T)(W)\delta, \end{aligned} \tag{3.25}$$

where $\delta = E[\ln(S)]$.

Solving for W we get

$$W = \frac{E[\ln(Z)] - E[\ln(X)]}{(1-T)\delta}. \tag{3.26}$$

Now the first term in the numerator, $E[\ln(Z)]$, can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\ln(Z_i)$$

and for the second term, $E[\ln(X)]$, we need to use the second order Taylor's approximation, whereby

$$\ln(X) \approx \ln(\mu_x) + (X - \mu_x)\frac{1}{\mu_x} - \frac{(X - \mu_x)^2}{2\mu_x^2}. \tag{3.27}$$

Then taking the expected values on both sides we get

$$E[\ln(X)] \approx \ln(\mu_x) - \frac{1}{2}\frac{Var(X)}{\mu_x^2}. \tag{3.28}$$

For μ_x we can use the estimator $\widehat{\mu}_x = \frac{\sum_{i=1}^{n} Z_i}{n}$. Also $Var(X)$, can be approximated by $Var(Z)$ since

$$E(Z) = E(X), \tag{3.29}$$

and using equation (3.20), it can be verified that

$$\begin{aligned} E(Z^2) &= E(X^2)[1 - (1 - T)W\{1 - E(S^2)\}] \\ &\approx E(X^2). \end{aligned} \tag{3.30}$$

Note that $(1-T)W$ is expected to be small, being product of two fractions, especially for large values of T. Now equation (3.28) becomes

$$E[\ln(X)] \approx \ln(\mu_x) - \frac{1}{2}\frac{Var(Z)}{\mu_x^2}. \tag{3.31}$$

Finally substituting equation (3.31) into equation (3.26) leads to an estimator of W given by

$$\widehat{W}_{G2} \approx \frac{\frac{1}{n}\sum_{i=1}^{n}\ln(Z_i) - \ln\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) + \frac{\widehat{V}(Z)}{2(\overline{Z})^2}}{(1 - T)\delta}. \tag{3.32}$$

Because of the term $(1 - T)$ in the denominator, the estimate for sensitivity level (W) in the two-stage version may be larger than the one-stage version, but there will be a gain in estimating the mean of the sensitive question (μ_x) since some of the respondents are forced to tell the truth

CHAPTER IV

PROPOSED TWO-STAGE OPTIONAL RRT MODELS

As indicated in Chapter 3, the two-stage optional RRT models proposed so far are not completely functional. The Ryu et al. (2006) model is not a true two-stage optional RRT model. The Gupta and Shabbir (2007) two-stage multiplicative optional RRT model using a single sample has an approximation drawback in estimating the sensitivity level W. In the current chapter, we try to address these issues.

4.1 Proposed Model 1: Two-stage Multiplicative Optional RRT Model using Two Independent Samples

4.1.1 Model Framework

Updating the two-stage multiplicative optional RRT model using a single sample to using two independent samples will at least give us an exact expression for W. As with the previous two-stage optional RRT model, a certain preset proportion (T) of respondents is instructed to answer truthfully (stage 1). The rest are instructed to answer truthfully if they consider the question to be non-sensitive, otherwise *multiplicatively* scramble their response using a scrambling device (stage 2.) But with this model, we will employ two different independent samples using two randomization devices with different distributions. This will allow us to get estimates for both μ_x and W simultaneously.

Thus specifying the mathematical model, we let X be the sensitive variable with *unknown* mean μ_x and *unknown* variance σ_x^2, and let W be the sensitivity

level. We select two independent random samples, of sizes n_i $(i = 1, 2)$, where ith sample uses randomization device R_i. The two randomization devices have different probability distributions with means θ_i and variances σ^2_{Si}. In each sample, a certain proportion (T) of the respondents is asked to answer truthfully (stage 1); and the rest asked to multiply the true response with the output of the randomization device and report the *product* as the final response.

Under this model the reported response Z_i $(i = 1, 2)$ for the ith sample, would be

$$Z_i = \begin{cases} X & \text{with probability } T + (1-T)(1-W) \\ S_i X & \text{with probability } (1-T)W \end{cases}$$

where X is the true response, S_i $(i = 1, 2)$ is the scrambling variable and W is the sensitivity level. We assume X, S_1,and S_2 to be mutually independent.

Assuming $E(X) = \mu_x$ and $E(S_i) = \theta_i$, one can calculate the expected value of the reported response (Z_i) as shown below.

$$\begin{aligned} E(Z_i) &= E(X)[T + (1-T)(1-W)] + E(S_i)E(X)[(1-T)W] \\ &= \mu_x[T + (1-T)(1-W)] + \theta_i \mu_x (1-T)W \\ &= \mu_x[T + (1-T)(1-W) + \theta_i(1-T)W] \\ &= \mu_x[(\theta_i W(1-T) + (1 - W + WT)] \\ &= \mu_x[(\theta_i W(1-T) + (1 - W(1-T)], \qquad i = 1, 2. \end{aligned} \tag{4.1}$$

Thus our two expected values for Z_1 and Z_2 are

$$E(Z_1) = \mu_x[(\theta_1 W(1-T) + (1 - W(1-T)] \tag{4.2}$$

and

$$E(Z_2) = \mu_x[(\theta_2 W(1-T) + (1 - W(1-T)]. \tag{4.3}$$

Solving simultaneously for W and μ_x we get

$$\mu_x = \frac{E(Z_2)(\theta_1 - 1) - E(Z_1)(\theta_2 - 1)}{(\theta_1 - \theta_2)}, \quad \theta_1 \neq \theta_2, \tag{4.4}$$

and

$$W = \frac{E(Z_2) - E(Z_1)}{[E(Z_1)(\theta_2 - 1) - E(Z_2)(\theta_1 - 1)]\,(1 - T)}. \tag{4.5}$$

Now, estimating $E(Z_i)$ by $\overline{Z}_i$, we get

$$\widehat{\mu}_x = \frac{\overline{Z}_2(\theta_1 - 1) - \overline{Z}_1(\theta_2 - 1)}{(\theta_1 - \theta_2)}, \quad \theta_1 \neq \theta_2, \tag{4.6}$$

and

$$\widehat{W} = \frac{\overline{Z}_2 - \overline{Z}_1}{\left[\overline{Z}_1(\theta_2 - 1) - \overline{Z}_2(\theta_1 - 1)\right](1 - T)}. \tag{4.7}$$

While this approach provides us an exact estimate of W (unlike the Gupta and Shabbir (2007) model) it still represents $\widehat{W}$ as a ratio of two random variables. Hence some approximation will be needed in order to define variance properties of $\widehat{W}$. Hence we decided to not pursue this approach any further.

4.2 Proposed Model 2: Two-stage Additive Optional RRT Model using Two Independent Samples

4.2.1 Model Framework

To eliminate the approximation problems involved in deriving expressions for $\widehat{W}$ and $Var(\widehat{W})$ using the two-stage multiplicative optional RRT models (one sample or two sample versions,) we propose an *additive* version of the two-stage optional RRT model using two independent samples. This model allows for exact expressions for estimators, $\widehat{\mu}_x$ and $\widehat{W}$, and their variances, as well as providing an added bonus of making it much more user friendly for the respondents.

As with the multiplicative model described above, there are "two stages". Rather than having two separate decks of cards for each stage, both stages are specified within the same deck. This reduces the respondent and interviewer burden, and as a practical matter, makes it much easier to handle in an experimental environment. Essentially, the randomization device (deck of cards) contains a certain known proportion of "truth" cards (this represents stage 1) and the rest of the cards list some numbers on them (this represents stage 2.) These numbers follow a known probability distribution. If the respondent picks the "truth" card, he is asked to answer the sensitive question truthfully. But if the respondent picks a numeric card, he is asked to answer truthfully if he does not believe the question is "sensitive" If the respondent does believe the question is sensitive, he is asked to give a scrambled response, by adding the number listed on the card picked to the true answer, and reporting the final response. As usual, the interviewer does not know if the respondent answered truthfully in stage one or, if relegated to stage two, then provided a scrambled response (based on his own sensitivity classification.) The interviewer simply records *one* numeric response.

4.2.2 Specific Model using two independent samples

Since there are two parameters that need estimation, two independent random samples with sample sizes n_1 and n_2 are employed. Sample i uses the randomization device R_i $(i = 1, 2)$. In each sample, a proportion (T) of respondents is instructed to answer truthfully. The rest of the subjects in sample i use randomization device R_i and provide an optionally scrambled response using the additive model. The two randomization devices have known means θ_i and known variances σ_i^2.

As usual, X is the sensitive question variable with *unknown* mean μ_x and *unknown* variance σ_x^2. S_i $(i = 1, 2)$ is the scrambling variable corresponding to the randomization device R_i $(i = 1, 2)$ and W is the sensitivity level. We assume X, S_1, and S_2 to be mutually independent.

Under this model, the reported response Z_i $(i = 1, 2)$ for the ith sample, would be

$$Z_i = \begin{cases} X & \text{with probability } T + (1-T)(1-W) \\ X + S_i & \text{with probability } (1-T)W. \end{cases}$$

Note that

$$\begin{aligned} E(Z_i) &= E(X)[T + (1-T)(1-W)] + E(X+S_i)[(1-T)(W)] \\ &= E(X)T + E(X)(1-T)(1-W) \\ &\quad + E(X)(1-T)(W) + E(S_i)(1-T)(W) \\ &= E(X)T + E(X)(1-T) + E(S_i)(1-T)(W) \\ &\quad - E(X)(1-T)W + E(X)(1-T)(W) \\ &= E(X)T - E(X)(T) + E(X) + E(S_i)(1-T)(W) \\ &= \mu_x + \theta_i W(1-T), \qquad\qquad i = 1, 2. \end{aligned} \tag{4.8}$$

Similarly

$$E(Z_i^2) = E\left(X^2\right)[T + (1-T)(1-W)] + E\left[(X+S_i)^2\right][(1-T)(W)]$$

$$
\begin{aligned}
&= E\left(X^2\right)T + E\left(X^2\right)(1-T)(1-W) \\
&\quad + E\left(X^2 + 2XS_i + S_i^2\right)(1-T)(W) \\
&= E\left(X^2\right)T + E\left(X^2\right)(1-T)(1-W) + E(X^2)(1-T)(W) \\
&\quad + 2E(X)E(S_i)(1-T)(W) + E(S_i^2)(1-T)(W) \\
&= E\left(X^2\right) + \left[2E(X)E(S_i) + E(S_i^2)\right]W(1-T) \\
&= \sigma_x^2 + \mu_x^2 + (2\mu_x\theta_i + \sigma_{si}^2 + \theta_i^2)W(1-T), \qquad i = 1, 2. \qquad (4.9)
\end{aligned}
$$

Thus

$$
\begin{aligned}
\sigma_{zi}^2 &= E(Z_i^2) - [E(Z_i)]^2 \\
&= \sigma_x^2 + \mu_x^2 + (2\mu_x\theta_i + \sigma_{si}^2 + \theta_i^2)W(1-T) \\
&\quad - [\mu_x + \theta_i(W)(1-T)]^2 \\
&= \sigma_x^2 + \sigma_{Si}^2 W(1-T) + \theta_i^2 W(1-T)[1 - W(1-T)]. \qquad (4.10)
\end{aligned}
$$

Hence

$$
\sigma_{Z1}^2 = \sigma_x^2 + \sigma_{S1}^2 W(1-T) + \theta_1^2 W(1-T)[1 - W(1-T)] \qquad (4.11)
$$

and

$$
\sigma_{Z2}^2 = \sigma_x^2 + \sigma_{S2}^2 W(1-T) + \theta_2^2 W(1-T)[1 - W(1-T)], \qquad (4.12)
$$

where σ_x^2 is the variance of true response variable X and σ_{Si}^2 is the (known) variance of the ith scrambling device S_i $(i = 1, 2)$.

Solving (4.8) simultaneously for μ_x and W we get

$$
\mu_x = \frac{E(Z_1)\theta_2 - E(Z_2)\theta_1}{\theta_2 - \theta_1}, \qquad \theta_1 \neq \theta_2,
$$

and

$$
W = \frac{E(Z_2) - E(Z_1)}{(\theta_2 - \theta_1)(1-T)}, \qquad \theta_1 \neq \theta_2.
$$

Now, estimating $E(Z_i)$ by $\overline{Z}_i$ we get the following estimators for our proposed model:

$$\widehat{\mu}_x = \frac{\overline{Z}_1\theta_2 - \overline{Z}_2\theta_1}{\theta_2 - \theta_1}, \quad \theta_1 \neq \theta_2, \tag{4.13}$$

and

$$\widehat{W} = \frac{\overline{Z}_2 - \overline{Z}_1}{(\theta_2 - \theta_1)(1-T)}, \quad \theta_1 \neq \theta_2. \tag{4.14}$$

It is easily verified that these estimators ($\widehat{\mu}_x$ and $\widehat{W}$) are unbiased estimators of the true population mean μ_x and the true population sensitivity W, respectively, as shown below.

$$\begin{aligned}
E\left(\widehat{\mu}_x\right) &= E\left(\frac{\overline{Z}_1\theta_2 - \overline{Z}_2\theta_1}{\theta_2 - \theta_1}\right) \\
&= \frac{1}{(\theta_2 - \theta_1)} E(\overline{Z}_1\theta_2 - \overline{Z}_2\theta_1) \\
&= \frac{1}{(\theta_2 - \theta_1)} \left[E(\overline{Z}_1\theta_2) - E(\overline{Z}_2\theta_1)\right] \\
&= \frac{1}{(\theta_2 - \theta_1)} \left[\theta_2 E(\overline{Z}_1) - \theta_1 E(\overline{Z}_2)\right] \\
&= \frac{1}{(\theta_2 - \theta_1)} \left[\theta_2 E(Z_1) - \theta_1 E(Z_2)\right] \\
&= \frac{1}{(\theta_2 - \theta_1)} \left[\theta_2 \left\{\mu_x + \theta_1 W(1-T)\right\} - \theta_1 \left\{\mu_x + \theta_2 W(1-T)\right\}\right] \\
&= \frac{1}{(\theta_2 - \theta_1)} (\theta_2 - \theta_1)\mu_x \\
&= \mu_x,
\end{aligned}$$

and

$$E\left(\widehat{W}\right) = E\left(\frac{\overline{Z}_2 - \overline{Z}_1}{(\theta_2 - \theta_1)(1-T)}\right)$$

$$
\begin{aligned}
&= \frac{1}{(\theta_2-\theta_1)(1-T)}E(\overline{Z}_2-\overline{Z}_1)\\
&= \frac{1}{(\theta_2-\theta_1)(1-T)}[E(\overline{Z}_2)-E(\overline{Z}_1)]\\
&= \frac{1}{(\theta_2-\theta_1)(1-T)}[E(Z_2)-E(Z_1)]\\
&= \frac{1}{(\theta_2-\theta_1)(1-T)}[\{\mu_x+\theta_2W(1-T)\}-\{\mu_x+\theta_1W(1-T)\}]\\
&= \frac{1}{(\theta_2-\theta_1)(1-T)}[\mu_x+\theta_2W(1-T)-\mu_x-\theta_1W(1-T)]\\
&= \frac{1}{(\theta_2-\theta_1)(1-T)}(\theta_2-\theta_1)(1-T)W\\
&= W.
\end{aligned}
$$

Furthermore

$$
\begin{aligned}
Var\left(\widehat{\mu}_x\right) &= Var\left(\frac{\overline{Z}_1\theta_2-\overline{Z}_2\theta_1}{\theta_2-\theta_1}\right)\\
&= \frac{1}{(\theta_2-\theta_1)^2}Var(\overline{Z}_1\theta_2-\overline{Z}_2\theta_1)\\
&= \frac{1}{(\theta_2-\theta_1)^2}[\theta_2^2Var(\overline{Z}_1)+Var(\theta_1^2\overline{Z}_2)]\\
&= \frac{1}{(\theta_2-\theta_1)^2}\left[\theta_2^2\left(\frac{\sigma_{Z_1}^2}{n_1}\right)+\theta_1^2\left(\frac{\sigma_{Z_2}^2}{n_2}\right)\right]. \qquad (4.15)
\end{aligned}
$$

Similarly

$$
\begin{aligned}
Var\left(\widehat{W}\right) &= Var\left(\frac{\overline{Z}_2-\overline{Z}_1}{(\theta_2-\theta_1)(1-T)}\right)\\
&= \frac{1}{(\theta_2-\theta_1)^2(1-T)^2}Var(\overline{Z}_2-\overline{Z}_1)\\
&= \frac{1}{(\theta_2-\theta_1)^2(1-T)^2}\left[\frac{\sigma_{Z_1}^2}{n_1}+\frac{\sigma_{Z_2}^2}{n_2}\right]. \qquad (4.16)
\end{aligned}
$$

Note that $\sigma^2_{Z_i}$ was calculated in (4.10). Both $Var\left(\widehat{\mu}_x\right)$ and $Var\left(\widehat{W}\right)$ can be estimated by their corresponding sample variances. The key advantage in using this approach is that we get exact expressions for our estimates and their means and variances, as compared to the previously introduced multiplicative model.

Using our previous example of "how many alcoholic drinks do you typically consume per week?" to demonstrate the procedural details of the two-stage additive optional RRT model, we could choose two independent subsamples of 5 members each. The two subsamples would be given *different* decks, but 20% ($T = .2$) of both decks would ask the respondents to answer truthfully. The remaining numeric portions of each of the two decks would have different scrambling distributions, with means (variances) $\theta_1 = 7$ ($\sigma^2_{S1} = 1.42$) and $\theta_2 = 3$ ($\sigma^2_{S1} = 1.92$), respectively. Suppose after collecting the experimental data we find that the sample means of the reported responses are as follows:

Sample 1	$\theta_1 = 7$, $\sigma^2_{S1} = 1.42$			
	$T = .2$	$(1 - T) = .8$		
True # of drinks	Stage 1	Stage 2	Sensitive?	Reported Response
5	go to S2	9	Y	14
1	go to S2	7	n	1
7	go to S2	5	Y	12
6	Tell Truth	N/A	(y)	6
3	go to S2	6	n	3
				$\overline{Z}_1 = 7.2$

Sample 2	$\theta_2 = 3$, $\sigma^2_{S2} = 1.92$			
	$T = .2$	$(1 - T) = .8$		
True # of drinks	Stage 1	Stage 2	Sensitive?	Reported Response
10	go to S2	4	n	10
2	Tell Truth	N/A	n	2
3	go to S2	5	n	3
7	go to S2	1	Y	8
5	go to S2	3	n	5
				$\overline{Z}_2 = 5.6$

Then we would calculate the estimates for μ_x and W using equations (4.13) and (4.14) as

$$\widehat{\mu}_x = \frac{\overline{Z}_1\theta_2 - \overline{Z}_2\theta_1}{\theta_2 - \theta_1} = \frac{7.2(3) - 5.6(7)}{(3-7)} = 4.4,$$

and

$$\widehat{W} = \frac{\overline{Z}_2 - \overline{Z}_1}{(\theta_2 - \theta_1)(1-T)} = \frac{5.6 - 7.2}{(3-7)(1-.2)} = .5.$$

Providing another example, now employing a larger sample size of $n = 100$, we may choose two independent subsamples of 50 members each and ask them to follow the two-stage additive optional model in answering a potentially sensitive question. The two subsamples would be given different decks, but 10% ($T = .1$) of both decks would ask the respondents to answer truthfully. The remaining numeric portions of each of the two decks would have different scrambling distributions, with means $\theta_1 = 2$ and $\theta_2 = 5$, respectively. Suppose after collecting the experimental data we find that the sample means of the reported responses are $\overline{Z}_1 = 4.46$ and $\overline{Z}_2 = 5.64$, respectively. The sample variances of Z_1 and Z_2 are 7.2739 and 10.7249, respectively. Using equations (4.13) and (4.14), we calculate the estimates for μ_x and W as

$$\widehat{\mu}_x = \frac{5(4.46) - 2(5.64)}{(5-2)} = 3.673,$$

and

$$\widehat{W} = \frac{(5.64 - 4.46)}{(5-2)(1-.1)} = .437.$$

Furthermore, using our sample variances of Z_1 and Z_2 as estimates of σ^2_{Z1} and σ^2_{Z2}, respectively, we use equations (4.15) and (4.16) to get

$$\widehat{V}ar(\widehat{\mu}_x) = \frac{1}{(5-2)^2}\left[25\left(\frac{7.2739}{50}\right) + 4\left(\frac{10.7249}{50}\right)\right] = .4994,$$

and

$$\widehat{Var}(\widehat{W}) = \frac{1}{(5-2)^2(1-.1)^2}\left[\frac{7.2739}{50} + \frac{10.7249}{50}\right] = .0444.$$

4.2.3 Sample Size Optimization

Although one can pick two independent samples n_1 and n_2 of equal sizes, this may have an effect of inflating the variances of estimators $\widehat{\mu}_x$ and $\widehat{W}$. Picking optimal combination of sample sizes can help minimize the variance. Thus, taking both variances into account, one can try to find n_1 and n_2 that minimize $[Var\left(\widehat{\mu}_x\right) + Var(\widehat{W})]$. We do this by taking partial derivatives with respect to n_1 and n_2, respectively, setting the derivatives to zero, then solving for n_1 and n_2 to find specific optimal sample sizes. The optimal sample sizes subject to $n_1 + n_2 = n$, are

$$n_1 = \frac{n\sigma_{Z_1}\sqrt{(1-T)^2\theta_2^2+1}}{\sigma_{Z_1}\sqrt{(1-T)^2\theta_2^2+1} + \sigma_{Z_2}\sqrt{(1-T)^2\theta_1^2+1}} \tag{4.17}$$

and

$$n_2 = \frac{n\sigma_{Z_2}\sqrt{(1-T)^2\theta_1^2+1}}{\sigma_{Z_1}\sqrt{(1-T)^2\theta_2^2+1} + \sigma_{Z_2}\sqrt{(1-T)^2\theta_1^2+1}}. \tag{4.18}$$

Table 4.1 below shows a comparison of variances $Var\left(\widehat{\mu}_x\right)$ and $Var(\widehat{W})$ when using equal sample sizes and optimal sample sizes. Note that $Var\left(\widehat{\mu}_x\right)$ decreases quite a bit when n_1 and n_2 are chosen optimally as compared to equal sample sizes, but $Var(\widehat{W})$ goes up slightly.

Table 4.1: Comparison of $V(\widehat{\mu}_x)$ and $V(\widehat{W})$ for equal and optimal sample sizes (optimized to minimize $\{V(\widehat{\mu}_x) + V(\widehat{W})\}$) for $\mu_x = 4$, $\sigma_x^2 = 4$, $\theta_1 = 2$, $\sigma_{S1}^2 = 2$, $\theta_2 = 5$, $\sigma_{S2}^2 = 5$, $T = .3$

		$n_1 = n_2$		optimum n_1, n_2			
n	W	$V(\widehat{\mu}_x)$	$V(\widehat{W})$	n_1	n_2	$V(\widehat{\mu}_x)$	$V(\widehat{W})$
100	0.1	0.2976	0.0471	64	36	0.2648	0.0532
	0.2	0.3331	0.0566	62	38	0.3035	0.0634
	0.3	0.3642	0.0648	61	39	0.3363	0.0724
	0.4	0.3909	0.0717	60	40	0.3644	0.0795
	0.5	0.4133	0.0773	60	40	0.3868	0.0860
	0.6	0.4314	0.0817	59	41	0.4059	0.0898
	0.7	0.4451	0.0847	59	41	0.4192	0.0932
	0.8	0.4544	0.0865	59	41	0.4279	0.0951
	0.9	0.4594	0.0869	59	41	0.4322	0.0956
	1	0.4600	0.0861	60	40	0.4306	0.0959
500	0.1	0.0595	0.0094	322	178	0.0529	0.0107
	0.2	0.0666	0.0113	312	188	0.0606	0.0128
	0.3	0.0728	0.0130	306	194	0.0672	0.0145
	0.4	0.0782	0.0143	301	199	0.0728	0.0159
	0.5	0.0827	0.0155	299	201	0.0774	0.0172
	0.6	0.0863	0.0163	297	203	0.0811	0.0181
	0.7	0.0890	0.0169	296	204	0.0838	0.0187
	0.8	0.0909	0.0173	296	204	0.0855	0.0191
	0.9	0.0919	0.0174	297	203	0.0863	0.0192
	1	0.0920	0.0172	298	202	0.0862	0.0191
1000	0.1	0.0298	0.0047	645	355	0.0264	0.0054
	0.2	0.0333	0.0057	624	376	0.0303	0.0064
	0.3	0.0364	0.0065	611	389	0.0336	0.0072
	0.4	0.0391	0.0072	603	397	0.0364	0.0080
	0.5	0.0413	0.0077	597	403	0.0387	0.0086
	0.6	0.0431	0.0082	594	406	0.0405	0.0090
	0.7	0.0445	0.0085	592	408	0.0419	0.0093
	0.8	0.0454	0.0086	592	408	0.0428	0.0095
	0.9	0.0459	0.0087	594	406	0.0432	0.0096
	1	0.0460	0.0086	597	403	0.0431	0.0095

CHAPTER V

SIMULATION STUDY AND CONCLUSIONS

In this chapter we present results of a simulation study and provide some concluding remarks. The simulations are carried out using SAS (Windows version 9.1). The SAS code is provided in Section 5.4 of this chapter. For all simulations, the sensitive variable X is chosen to be a Poisson variable with mean $\mu_x = 4$. Although one could use any distribution for the simulations, we have chosen the Poisson distribution because it has been used in other comparable studies, and also it is a reasonable distribution when dealing with rare behaviors, in the sense that not everyone engages in these behaviors. The scrambling variables are also chosen to be Poisson for similar reasons: comparability with other studies and agreement with the parent distribution of the sensitive variable X. The scrambling variables S_1 and S_2 are specified to have means $\theta_1 = 2$ and $\theta_2 = 5$, respectively. These particular values for the means have been used in previous studies, thus allowing for reasonable comparison of our proposed model with earlier models. The results are averaged over 1000 simulation runs each with a sample size of $n = 100$, $n = 500$, and $n = 1000$. Simulations using optimal sample sizes and equal sample sizes ($n_1 = n_2$) were run, as appropriate. The goals of these simulations are threefold. We want to verify that our simulated estimates agree with the correct values for μ_x and W. Similarly, we want to make sure that the simulated variances are also in agreement with our theoretical variances. Finally, we want to verify that as values of n, T, and/or W change, expected trends can be seen in simulated results.

5.1 Quantitative Additive Optional RRT Models with Optimal Sample Selection

We would like to point out that the partial RRT models generally have the smallest variance and the optional models (both one-stage and two-stage) have much higher values for $Var(\widehat{\mu}_x)$. However, one should keep in mind that the partial RRT models estimate only one parameter (μ_x), while the optional RRT models also estimate the sensitivity level (W). Hence, comparing a partial RRT model to an optional RRT model would be unreasonable. So our simulation study will focus on how well a two-stage optional RRT model performs in comparison to a one-stage optional RRT model.

Tables 5.1, 5.2 and 5.3 that follow show the true parameter values and the corresponding simulated values for the one-stage additive optional RRT model ($T = 0$) and two-stage additive optional RRT models ($T = .1$ and $T = .3$) *using optimum choices* of n_1 and n_2. Note that there is hardly any bias in the estimation of μ_x and W. For example, in Table 5.2 ($T = .1$), at $n = 500$, the bias in estimation of $\mu_x = 4$ ranges from a low of .00469 to a high of .01691. More specifically, note that $\widehat{\mu}_x$ is approximately $N[\mu_x, Var(\widehat{\mu}_x)]$ at large n_1 and n_2. Also, looking at an example in Table 5.2, at $n = 500$ and $W = .3$, notice that our estimate of μ_x based on 1000 simulations is 4.0125 and $\widehat{V}(\widehat{\mu}_x)$ is .07495. Then the 95% confidence interval for μ_x in this case would be $4.0125 \pm (1.96)(\sqrt{.07495}/\sqrt{1000}) = 4.0125 \pm .01699$. This overlaps our true $\mu_x = 4$. The same is true for all other scenarios of $\widehat{\mu}_x$ and $\widehat{W}$. This was clearly expected given that both $\widehat{\mu}_x$ and $\widehat{W}$ are unbiased. Also notice that for both $\widehat{\mu}_x$ and $\widehat{W}$, theoretical variances are in good agreement with simulated variances, even for small sample sizes. Another observation is that the estimation gets better (smaller variance) as the sample size increases. For example, in Table 5.2, for $W = .4$, $\widehat{V}(\widehat{\mu}_x) = .39563$ at $n = 100$, it drops to $\widehat{V}(\widehat{\mu}_x) = .07896$ at $n = 500$,

and further drops to $\widehat{V}(\widehat{\mu}_x) = .03760$ at $n = 1000$.

One can also note an interesting trend as W increases from $W = .1$ to $W = 1.0$. Note that for a given value of n and T, both $Var(\widehat{\mu}_x)$ and $Var(\widehat{W})$ increase until reaching some peak, and then start decreasing. This is because we have "optimized" the choices of n_1 and n_2 in such a manner as to minimize the sum of the two variances $(V(\widehat{\mu}_x) + V(\widehat{W}))$. This continual optimization of n_1 and n_2 has a significant impact on the variance calculations. In fact, the shift in variances from an increasing pattern to a decreasing pattern occurs at about the same time as the shift in n_1 from decreasing to increasing values. Thus, to do a better comparison of estimators, recreating these tables with equal $n's$ $(n_1 = n_2)$ would be more appropriate. This is what we will do in the next section.

Table 5.1: Simulation results for $T = 0$ (one stage additive optional RRT model) at various levels of n and W using optimum choices of n_1 and n_2 that minimize $\{V(\widehat{\mu}_x) + V(\widehat{W})\}$ for $\mu_x = 4$, $\sigma_x^2 = 4$, $\theta_1 = 2$, $\sigma_{S1}^2 = 2$, $\theta_2 = 5$, $\sigma_{S2}^2 = 5$

n	$n1$	$n2$	W	$\widehat{W}$	$\widehat{\mu}_x$	$V(\widehat{\mu}_x)$	$\widehat{V}(\widehat{\mu}_x)$	$V(\widehat{W})$	$\widehat{V}(\widehat{W})$
100	65	35	0.1	0.09264	4.02745	0.28059	0.27909	0.02922	0.02842
100	63	37	0.2	0.19055	4.03332	0.33033	0.33159	0.03592	0.03653
100	62	38	0.3	0.29276	4.02768	0.36946	0.37025	0.04118	0.04086
100	61	39	0.4	0.39582	4.02240	0.39905	0.40601	0.04468	0.04521
100	61	39	0.5	0.49914	3.99301	0.41852	0.38996	0.04725	0.04410
100	61	39	0.6	0.59768	3.99794	0.42866	0.39377	0.04826	0.04529
100	61	39	0.7	0.70046	3.99275	0.42945	0.41703	0.04769	0.04779
100	62	38	0.8	0.80334	3.98522	0.41992	0.41346	0.04627	0.04697
100	63	37	0.9	0.89739	4.00220	0.40073	0.39366	0.04315	0.04369
100	65	35	1	0.99113	4.01813	0.37070	0.37084	0.03883	0.03819
500	326	174	0.1	0.09595	4.01302	0.05610	0.05509	0.00586	0.00590
500	315	185	0.2	0.19620	4.01316	0.06607	0.06805	0.00718	0.00769
500	309	191	0.3	0.29566	4.01355	0.07392	0.07720	0.00821	0.00886
500	306	194	0.4	0.39471	4.01477	0.07978	0.08053	0.00896	0.00909
500	305	195	0.5	0.50038	4.00485	0.08370	0.08352	0.00945	0.00911
500	305	195	0.6	0.59933	4.00680	0.08573	0.08266	0.00965	0.00903
500	307	193	0.7	0.69915	4.00541	0.08582	0.08312	0.00960	0.00926
500	311	189	0.8	0.79892	4.00482	0.08395	0.08373	0.00928	0.00903
500	317	183	0.9	0.89811	4.00750	0.08009	0.07899	0.00869	0.00847
500	325	175	1	0.99403	4.01794	0.07414	0.07655	0.00777	0.00792
1000	652	348	0.1	0.09868	4.00459	0.02805	0.02811	0.00293	0.00279
1000	631	369	0.2	0.19626	4.00997	0.03303	0.03283	0.00360	0.00348
1000	619	381	0.3	0.29641	4.01002	0.03695	0.03614	0.00411	0.00399
1000	612	388	0.4	0.39659	4.00834	0.03989	0.03967	0.00448	0.00436
1000	610	390	0.5	0.50152	3.99816	0.04185	0.04026	0.00473	0.00452
1000	611	389	0.6	0.60150	3.99721	0.04286	0.04127	0.00483	0.00456
1000	615	385	0.7	0.70076	3.99855	0.04290	0.04131	0.00481	0.00467
1000	622	378	0.8	0.80037	3.99988	0.04198	0.03949	0.00464	0.00445
1000	633	367	0.9	0.89804	4.00488	0.04005	0.03726	0.00434	0.00411
1000	651	349	1	1.00147	3.99423	0.03706	0.03807	0.00389	0.00398

Table 5.2: Simulation results for $T = .1$ (two stage additive optional RRT model) at various levels of n and W using optimum choices of n_1 and n_2 that minimize $\{V(\widehat{\mu}_x) + V(\widehat{W})\}$ for $\mu_x = 4$, $\sigma_x^2 = 4$, $\theta_1 = 2$, $\sigma_{S1}^2 = 2$, $\theta_2 = 5$, $\sigma_{S2}^2 = 5$

n	$n1$	$n2$	W	$\widehat{W}$	$\widehat{\mu}_x$	$V(\widehat{\mu}_x)$	$\widehat{V}(\widehat{\mu}_x)$	$V(\widehat{W})$	$\widehat{V}(\widehat{W})$
100	65	35	0.1	0.09076	4.02909	0.27514	0.27429	0.03498	0.03389
100	63	37	0.2	0.18754	4.03548	0.32145	0.32411	0.04263	0.04325
100	62	38	0.3	0.29087	4.03358	0.35893	0.36049	0.04889	0.04743
100	61	39	0.4	0.39462	4.02384	0.38864	0.39563	0.05335	0.05371
100	61	39	0.5	0.49265	4.02822	0.40995	0.41574	0.05699	0.05777
100	61	39	0.6	0.59940	3.99356	0.42370	0.39887	0.05907	0.05593
100	61	39	0.7	0.69657	3.99825	0.42988	0.39911	0.05957	0.05736
100	61	39	0.8	0.80103	3.98799	0.42849	0.40625	0.05851	0.05655
100	62	38	0.9	0.90455	3.98354	0.41852	0.40787	0.05674	0.05726
100	63	37	1	0.99710	4.00220	0.40073	0.39366	0.05327	0.05393
500	325	175	0.1	0.09442	4.01529	0.05503	0.05416	0.00700	0.00702
500	315	185	0.2	0.19596	4.01272	0.06429	0.06408	0.00853	0.00890
500	309	191	0.3	0.29588	4.01250	0.07182	0.07495	0.00975	0.01054
500	305	195	0.4	0.39328	4.01691	0.07773	0.07896	0.01067	0.01111
500	303	197	0.5	0.49317	4.01633	0.08206	0.08155	0.01133	0.01135
500	303	197	0.6	0.59890	4.00778	0.08481	0.08389	0.01174	0.01126
500	304	196	0.7	0.69886	4.00638	0.08601	0.08425	0.01188	0.01134
500	306	194	0.8	0.79823	4.00622	0.08566	0.08251	0.01174	0.01121
500	309	191	0.9	0.89873	4.00469	0.08374	0.08420	0.01131	0.01111
500	314	186	1	0.99845	4.00615	0.08018	0.08006	0.01062	0.01050
1000	651	349	0.1	0.09792	4.00602	0.02751	0.02755	0.00350	0.00334
1000	630	370	0.2	0.19619	4.00935	0.03215	0.03158	0.00426	0.00402
1000	617	383	0.3	0.29598	4.01011	0.03592	0.03553	0.00487	0.00473
1000	610	390	0.4	0.39627	4.00833	0.03886	0.03760	0.00533	0.00502
1000	606	394	0.5	0.49768	4.00326	0.04103	0.04055	0.00566	0.00549
1000	606	394	0.6	0.60130	3.99896	0.04240	0.04100	0.00587	0.00556
1000	607	393	0.7	0.70176	3.99604	0.04301	0.04228	0.00593	0.00576
1000	611	389	0.8	0.80140	3.99641	0.04284	0.04124	0.00586	0.00573
1000	618	382	0.9	0.90022	4.00018	0.04187	0.03950	0.00566	0.00539
1000	629	371	1	0.99844	4.00333	0.04008	0.03766	0.00532	0.00508

Table 5.3: Simulation results for $T = .3$ (two stage additive optional RRT model) at various levels of n and W using optimum choices of n_1 and n_2 that minimize $\{V(\widehat{\mu}_x) + V(\widehat{W})\}$ for $\mu_x = 4$, $\sigma_x^2 = 4$, $\theta_1 = 2$, $\sigma_{S1}^2 = 2$, $\theta_2 = 5$, $\sigma_{S2}^2 = 5$

n	$n1$	$n2$	W	$\widehat{W}$	$\widehat{\mu}_x$	$V(\widehat{\mu}_x)$	$\widehat{V}(\widehat{\mu}_x)$	$V(\widehat{W})$	$\widehat{V}(\widehat{W})$
100	64	36	0.1	0.09075	4.02284	0.26479	0.25858	0.05324	0.05037
100	62	38	0.2	0.18747	4.02654	0.30351	0.30902	0.06342	0.06287
100	61	39	0.3	0.28603	4.03424	0.33631	0.34433	0.07237	0.07522
100	60	40	0.4	0.38856	4.03447	0.36444	0.36952	0.07946	0.08043
100	60	40	0.5	0.49173	4.02614	0.38681	0.40040	0.08604	0.08995
100	59	41	0.6	0.59059	4.03085	0.40589	0.41913	0.08977	0.09244
100	59	41	0.7	0.68601	4.03728	0.41917	0.43281	0.09321	0.09604
100	59	41	0.8	0.79560	4.00127	0.42795	0.41528	0.09514	0.09290
100	59	41	0.9	0.89180	4.00665	0.43222	0.41713	0.09557	0.09536
100	60	40	1	0.99782	3.99889	0.43056	0.42338	0.09586	0.09707
500	322	178	0.1	0.09462	4.01136	0.05289	0.05174	0.01071	0.01060
500	312	188	0.2	0.19407	4.01388	0.06062	0.05922	0.01276	0.01293
500	306	194	0.3	0.29524	4.01184	0.06722	0.06842	0.01452	0.01521
500	301	199	0.4	0.39447	4.01406	0.07284	0.07394	0.01594	0.01670
500	299	201	0.5	0.49331	4.01442	0.07741	0.07798	0.01716	0.01774
500	297	203	0.6	0.59285	4.01445	0.08107	0.07900	0.01806	0.01781
500	296	204	0.7	0.69116	4.01653	0.08378	0.07941	0.01869	0.01797
500	296	204	0.8	0.79823	4.00898	0.08553	0.08492	0.01908	0.01864
500	297	203	0.9	0.89806	4.00740	0.08633	0.08498	0.01922	0.01850
500	298	202	1	0.99752	4.00793	0.08622	0.08234	0.01906	0.01819
1000	645	355	0.1	0.09763	4.00507	0.02643	0.02708	0.00537	0.00526
1000	624	376	0.2	0.19684	4.00682	0.03031	0.03036	0.00638	0.00614
1000	611	389	0.3	0.29507	4.00911	0.03362	0.03316	0.00725	0.00692
1000	603	397	0.4	0.39594	4.00810	0.03641	0.03611	0.00798	0.00782
1000	597	403	0.5	0.49466	4.01000	0.03872	0.03777	0.00857	0.00820
1000	594	406	0.6	0.59487	4.00868	0.04053	0.04054	0.00903	0.00896
1000	592	408	0.7	0.69451	4.00846	0.04189	0.04188	0.00935	0.00918
1000	592	408	0.8	0.80113	4.00034	0.04277	0.04120	0.00954	0.00906
1000	594	406	0.9	0.90208	3.99661	0.04316	0.04204	0.00961	0.00935
1000	597	403	1	1.00252	3.99563	0.04310	0.04196	0.00954	0.00949

5.2 Simulation Results based on Equal Sample Sizes

Similar to Section 5.1, the following three Tables (5.4, 5.5 and 5.6) show the true parameter values and corresponding simulated values for the one-stage additive optional RRT model ($T = 0$) and two-stage additive optional RRT models ($T = .1$ and $T = .3$) but this time *using equal sample sizes* ($n_1 = n_2$).

Once again we observe that the simulated values are all in good agreement with theoretical values. For example, in Table 5.5, for $n = 500$ and $W = .2$, the true $V(\widehat{\mu}_x) = .07028$, while the simulated $\widehat{V}(\widehat{\mu}_x) = .07052$. We also note, as in the case of optimal sample sizes, that both $Var(\widehat{\mu}_x)$ and $Var(\widehat{W})$ increase as W increases, but only up to a certain value of W. After that, these values begin to drop. Note that $Var(\widehat{\mu}_x)$ and $Var(\widehat{W})$ peak for different values of W. On further examination, an explanation can be identified. It is presented after Table 5.6.

Table 5.4: Simulation results for $T = 0$ (one stage additive optional RRT model) at various levels of n and W using equal n_1 and n_2 with $\mu_x = 4$, $\sigma_x^2 = 4$, $\theta_1 = 2$, $\sigma_{S1}^2 = 2$, $\theta_2 = 5$, $\sigma_{S2}^2 = 5$

n	$n1$	$n2$	W	$\widehat{W}$	$\widehat{\mu}_x$	$V(\widehat{\mu}_x)$	$\widehat{V}(\widehat{\mu}_x)$	$V(\widehat{W})$	$\widehat{V}(\widehat{W})$
100	50	50	0.1	0.09018	4.03966	0.31333	0.31203	0.02513	0.02478
100	50	50	0.2	0.18740	4.04798	0.36000	0.35814	0.03120	0.03155
100	50	50	0.3	0.28839	4.04603	0.39778	0.39283	0.03598	0.03571
100	50	50	0.4	0.38764	4.05074	0.42667	0.41028	0.03947	0.03787
100	50	50	0.5	0.49669	4.00273	0.44667	0.43499	0.04167	0.04055
100	50	50	0.6	0.59471	4.01015	0.45778	0.45201	0.04258	0.04208
100	50	50	0.7	0.69395	4.01485	0.46000	0.45704	0.04220	0.04212
100	50	50	0.8	0.79495	4.01291	0.45333	0.45536	0.04053	0.04097
100	50	50	0.9	0.89217	4.02123	0.43778	0.44233	0.03758	0.03822
100	50	50	1	0.99209	4.01937	0.41333	0.42782	0.03333	0.03414
500	250	250	0.1	0.09688	4.01302	0.06267	0.06182	0.00503	0.00490
500	250	250	0.2	0.19700	4.01303	0.07200	0.07414	0.00624	0.00650
500	250	250	0.3	0.29696	4.01181	0.07956	0.08060	0.00720	0.00729
500	250	250	0.4	0.39625	4.01206	0.08533	0.08462	0.00789	0.00765
500	250	250	0.5	0.49818	4.01210	0.08933	0.09532	0.00833	0.00854
500	250	250	0.6	0.59668	4.01561	0.09156	0.09125	0.00852	0.00821
500	250	250	0.7	0.69597	4.01586	0.09200	0.09054	0.00844	0.00816
500	250	250	0.8	0.79593	4.01464	0.09067	0.09278	0.00811	0.00801
500	250	250	0.9	0.89605	4.01465	0.08756	0.09041	0.00752	0.00767
500	250	250	1	0.99587	4.01348	0.08267	0.07801	0.00667	0.00616
1000	500	500	0.1	0.09883	4.00443	0.03133	0.03115	0.00251	0.00234
1000	500	500	0.2	0.19794	4.00581	0.03600	0.03437	0.00312	0.00274
1000	500	500	0.3	0.29752	4.00713	0.03978	0.03723	0.00360	0.00321
1000	500	500	0.4	0.39815	4.00421	0.04267	0.04128	0.00395	0.00363
1000	500	500	0.5	0.49996	4.00258	0.04467	0.04488	0.00417	0.00424
1000	500	500	0.6	0.60032	4.00047	0.04578	0.04450	0.00426	0.00411
1000	500	500	0.7	0.70094	3.99755	0.04600	0.04613	0.00422	0.00428
1000	500	500	0.8	0.80053	3.99887	0.04533	0.04503	0.00405	0.00415
1000	500	500	0.9	0.89964	4.00025	0.04378	0.04196	0.00376	0.00368
1000	500	500	1	1.00405	3.98547	0.04133	0.03952	0.00333	0.00322

Table 5.5: Simulation results for $T = .1$ (two stage additive optional RRT model) at various levels of n and W using equal n_1 and n_2 with $\mu_x = 4$, $\sigma_x^2 = 4$, $\theta_1 = 2$, $\sigma_{S1}^2 = 2$, $\theta_2 = 5$, $\sigma_{S2}^2 = 5$

n	$n1$	$n2$	W	$\widehat{W}$	$\widehat{\mu}_x$	$V(\widehat{\mu}_x)$	$\widehat{V}(\widehat{\mu}_x)$	$V(\widehat{W})$	$\widehat{V}(\widehat{W})$
100	50	50	0.1	0.08908	4.03871	0.30818	0.30998	0.03019	0.02989
100	50	50	0.2	0.18310	4.05299	0.35138	0.35210	0.03715	0.03746
100	50	50	0.3	0.28611	4.05210	0.38738	0.38299	0.04281	0.04181
100	50	50	0.4	0.38697	4.04851	0.41618	0.40827	0.04719	0.04610
100	50	50	0.5	0.48501	4.05369	0.43778	0.43458	0.05028	0.05012
100	50	50	0.6	0.59470	4.00865	0.45218	0.45009	0.05208	0.05216
100	50	50	0.7	0.69173	4.01467	0.45938	0.45449	0.05259	0.05248
100	50	50	0.8	0.79327	4.01125	0.45938	0.45328	0.05181	0.05093
100	50	50	0.9	0.89584	4.00930	0.45218	0.44900	0.04975	0.04995
100	50	50	1	0.99130	4.02123	0.43778	0.44233	0.04639	0.04719
500	250	250	0.1	0.09576	4.01446	0.06164	0.06035	0.00604	0.00583
500	250	250	0.2	0.19577	4.01540	0.07028	0.07052	0.00743	0.00759
500	250	250	0.3	0.29731	4.01070	0.07748	0.07915	0.00856	0.00878
500	250	250	0.4	0.39522	4.01354	0.08324	0.08257	0.00944	0.00923
500	250	250	0.5	0.49550	4.01191	0.08756	0.08618	0.01006	0.00971
500	250	250	0.6	0.59594	4.01667	0.09044	0.09354	0.01042	0.01038
500	250	250	0.7	0.69591	4.01535	0.09188	0.09201	0.01052	0.01011
500	250	250	0.8	0.79472	4.01652	0.09188	0.09037	0.01036	0.00995
500	250	250	0.9	0.89583	4.01339	0.09044	0.09309	0.00995	0.00994
500	250	250	1	0.99561	4.01465	0.08756	0.09041	0.00928	0.00947
1000	500	500	0.1	0.09875	4.00427	0.03082	0.03077	0.00302	0.00283
1000	500	500	0.2	0.19791	4.00554	0.03514	0.03371	0.00371	0.00325
1000	500	500	0.3	0.29737	4.00689	0.03874	0.03693	0.00428	0.00382
1000	500	500	0.4	0.39773	4.00492	0.04162	0.03941	0.00472	0.00423
1000	500	500	0.5	0.49902	3.99983	0.04378	0.04193	0.00503	0.00456
1000	500	500	0.6	0.59926	4.00453	0.04522	0.04515	0.00521	0.00517
1000	500	500	0.7	0.70017	4.00009	0.04594	0.04573	0.00526	0.00521
1000	500	500	0.8	0.80083	3.99750	0.04594	0.04531	0.00518	0.00523
1000	500	500	0.9	0.90046	3.99892	0.04522	0.04509	0.00497	0.00508
1000	500	500	1	0.99960	4.00025	0.04378	0.04196	0.00464	0.00454

Table 5.6: Simulation results for $T = .3$ (two stage additive optional RRT model) at various levels of n and W using equal n_1 and n_2 with $\mu_x = 4$, $\sigma_x^2 = 4$, $\theta_1 = 2$, $\sigma_{S1}^2 = 2$, $\theta_2 = 5$, $\sigma_{S2}^2 = 5$

n	$n1$	$n2$	W	$\widehat{W}$	$\widehat{\mu}_x$	$V(\widehat{\mu}_x)$	$\widehat{V}(\widehat{\mu}_x)$	$V(\widehat{W})$	$\widehat{V}(\widehat{W})$
100	50	50	0.1	0.08530	4.03893	0.29760	0.29768	0.04707	0.04641
100	50	50	0.2	0.17981	4.04799	0.33307	0.33162	0.05656	0.05599
100	50	50	0.3	0.28069	4.05072	0.36418	0.36641	0.06477	0.06633
100	50	50	0.4	0.38343	4.04964	0.39093	0.38377	0.07168	0.06952
100	50	50	0.5	0.48390	4.04645	0.41333	0.40860	0.07731	0.07691
100	50	50	0.6	0.58369	4.04918	0.43138	0.42506	0.08165	0.08008
100	50	50	0.7	0.67859	4.05747	0.44507	0.44233	0.08470	0.08440
100	50	50	0.8	0.79187	4.01109	0.45440	0.45206	0.08647	0.08602
100	50	50	0.9	0.88936	4.01467	0.45938	0.45449	0.08694	0.08675
100	50	50	1	0.99136	4.01485	0.46000	0.45704	0.08612	0.08596
500	250	250	0.1	0.09600	4.01088	0.05952	0.05867	0.00941	0.00909
500	250	250	0.2	0.19511	4.01404	0.06661	0.06500	0.01131	0.01110
500	250	250	0.3	0.29655	4.01077	0.07284	0.07511	0.01295	0.01354
500	250	250	0.4	0.39668	4.01115	0.07819	0.07957	0.01434	0.01462
500	250	250	0.5	0.49482	4.01289	0.08267	0.08332	0.01546	0.01543
500	250	250	0.6	0.59511	4.01115	0.08628	0.08518	0.01633	0.01572
500	250	250	0.7	0.69280	4.01478	0.08901	0.08647	0.01694	0.01625
500	250	250	0.8	0.79516	4.01654	0.09088	0.09192	0.01729	0.01704
500	250	250	0.9	0.89474	4.01535	0.09188	0.09201	0.01739	0.01671
500	250	250	1	0.99425	4.01586	0.09200	0.09054	0.01722	0.01666
1000	500	500	0.1	0.09850	4.00353	0.02976	0.03022	0.00471	0.00451
1000	500	500	0.2	0.19913	4.00231	0.03331	0.03283	0.00566	0.00524
1000	500	500	0.3	0.29664	4.00663	0.03642	0.03457	0.00648	0.00565
1000	500	500	0.4	0.39674	4.00718	0.03909	0.03735	0.00717	0.00645
1000	500	500	0.5	0.49679	4.00620	0.04133	0.03912	0.00773	0.00693
1000	500	500	0.6	0.59757	4.00343	0.04314	0.04201	0.00817	0.00764
1000	500	500	0.7	0.69756	4.00241	0.04451	0.04325	0.00847	0.00784
1000	500	500	0.8	0.79916	4.00429	0.04544	0.04494	0.00865	0.00845
1000	500	500	0.9	0.90022	4.00009	0.04594	0.04573	0.00869	0.00862
1000	500	500	1	1.00135	3.99755	0.04600	0.04613	0.00861	0.00874

Using equations (4.11) and (4.12), $Var\left(\widehat{\mu}_x\right)$, as given in equation (4.15), can be written as

$$\begin{aligned} Var\left(\widehat{\mu}_x\right) &= \frac{1}{(\theta_2-\theta_1)^2}\left[\theta_2^2\left(\frac{\sigma_{Z_1}^2}{n_1}\right)+\theta_1^2\left(\frac{\sigma_{Z_2}^2}{n_2}\right)\right] \\ &= \frac{1}{(\theta_2-\theta_1)^2}\left[\theta_2^2\left(\frac{\sigma_x^2+\sigma_{S1}^2W(1-T)+\theta_1^2W(1-T)[1-W(1-T)]}{n_1}\right)\right. \\ &\qquad \left.+\theta_1^2\left(\frac{\sigma_x^2+\sigma_{S2}^2W(1-T)+\theta_2^2W(1-T)[1-W(1-T)]}{n_2}\right)\right]. \end{aligned}$$

If we let $\lambda = W(1-T)$ and $k = \frac{1}{(\theta_2-\theta_1)^2}$, we get

$$\begin{aligned} Var\left(\widehat{\mu}_x\right) &= k\left[\theta_2^2\left(\frac{\sigma_x^2+\sigma_{S1}^2\lambda+\theta_1^2\lambda[1-\lambda]}{n_1}\right)+\theta_1^2\left(\frac{\sigma_x^2+\sigma_{S2}^2\lambda+\theta_2^2\lambda[1-\lambda]}{n_2}\right)\right] \\ &= k\left[\frac{\theta_2^2}{n_1}(\sigma_x^2+\sigma_{S1}^2\lambda+\theta_1^2\lambda-\theta_1^2\lambda^2)+\frac{\theta_1^2}{n_2}(\sigma_x^2+\sigma_{S2}^2\lambda+\theta_2^2\lambda-\theta_2^2\lambda^2)\right] \\ &= k\left[\left(\frac{\theta_2^2}{n_1}+\frac{\theta_1^2}{n_2}\right)\sigma_x^2+\left(\frac{\theta_2^2}{n_1}(\sigma_{S1}^2+\theta_1^2)+\frac{\theta_1^2}{n_2}(\sigma_{S2}^2+\theta_2^2)\right)\lambda\right. \\ &\qquad \left.-\left(\frac{\theta_2^2\theta_1^2}{n_1}+\frac{\theta_1^2\theta_2^2}{n_2}\right)\lambda^2\right] \\ &= k\left\{c+\left[\frac{\theta_2^2}{n_1}(\sigma_{S1}^2+\theta_1^2)+\frac{\theta_1^2}{n_2}(\sigma_{S2}^2+\theta_2^2)\right]\lambda-\left[\frac{\theta_2^2\theta_1^2}{n_1}+\frac{\theta_1^2\theta_2^2}{n_2}\right]\lambda^2\right\} \end{aligned}$$

where $c = \left(\frac{\theta_2^2}{n_1}+\frac{\theta_1^2}{n_2}\right)\sigma_x^2$.

It can be verified that

$$W = \frac{\frac{\theta_2^2(\sigma_{S1}^2+\theta_1^2)}{n_1}+\frac{\theta_1^2(\sigma_{S2}^2+\theta_2^2)}{n_2}}{2\theta_2^2\theta_1^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)(1-T)} \tag{5.1}$$

is a point of maxima for $Var\left(\widehat{\mu}_x\right)$. If we let $n_1 = n_2$, the maximal point is given by

$$W = \frac{2\theta_1^2\theta_2^2+\sigma_{S1}^2\theta_2^2+\sigma_{S2}^2\theta_1^2}{4\theta_1^2\theta_2^2(1-T)}. \tag{5.2}$$

For our simulations, where, $\theta_1 = 2$, $\sigma^2_{S1} = 2$, $\theta_2 = 5$, and $\sigma^2_{S2} = 5$, we can expect to see the maximum $Var\left(\widehat{\mu}_x\right)$ at $W = \frac{.675}{(1-T)}$. For example, for $T = .3$, we should see the maximum for $Var\left(\widehat{\mu}_x\right)$ at $W = .675/.7 = .964$ and this is confirmed by looking at Table 5.6 above (the maximum is at about $W = 1.0$). Similarly, we should see the maximum for $T = 0$ at $W = .675$ and for $T = .1$ at $W = .75$.

We can similarly find the value of W for which $Var\left(\widehat{W}\right)$, becomes maximum for a given value of T. This point of maxima is given by

$$W = \frac{\frac{(\sigma^2_{S1}+\theta^2_1)}{n_1} + \frac{(\sigma^2_{S2}+\theta^2_2)}{n_2}}{2\left(\frac{\theta^2_1}{n_1} + \frac{\theta^2_2}{n_2}\right)(1-T)}. \tag{5.3}$$

If we let $n_1 = n_2$, then $Var\left(\widehat{W}\right)$ is maximum at

$$W = \frac{(\sigma^2_{S1} + \theta^2_1 + \sigma^2_{S2} + \theta^2_2)}{2(\theta^2_1 + \theta^2_2)(1-T)}. \tag{5.4}$$

So for our simulations, where, $\theta_1 = 2$, $\sigma^2_{S1} = 2$, $\theta_2 = 5$, and $\sigma^2_{S2} = 5$, we can expect to see the maximum $Var\left(\widehat{W}\right)$ at $W = \frac{.621}{(1-T)}$. For example, for $T = .3$, we should see the maximum for $Var\left(\widehat{W}\right)$ at $W = .621/.7 = .887$ and this is confirmed by looking at Table 5.6 above (the maximum is at about $W = .9$). Similarly, we should see the maximum for $T = 0$ at $W = .621$ and for $T = .1$ at $W = .69$.

We would like to point out one other aspect of the comparison of a one-stage optional RRT model ($T = 0$) with a two-stage optional RRT model ($T > 0$). Tables 5.7 through 5.10 take a closer look at how $Var\left(\widehat{\mu}_x\right)$ and $Var(\widehat{W})$ change as T increases from $T = 0$ (aka the one-stage optional model) to $T = .9$ for different values of W. Tables 5.7 and 5.8 use optimal sample sizes, while Tables 5.9 and 5.10 use equal sample sizes. For any fixed value of W, we expect $Var\left(\widehat{\mu}_x\right)$ to decrease and $Var(\widehat{W})$ to increase as T increases. This is because as more and

more respondents tell the truth, the estimation of μ_x should improve. At the same time, as more respondents tell the truth, there is a smaller pool of respondents that scramble their responses, and hence $Var(\widehat{W})$ increases. However, this trend is valid only up to moderate values of W. One should note that the penalty for using any RRT model is dependent on the sample size involved. So as T increases, on one hand there is a "gain", since more respondents are providing truthful responses, but on the other hand there is a "loss" since a smaller sample size is involved in the RRT part, thus increasing the penalty. For a two-stage model to be better than a one-stage model, the "gain" should be greater than the "loss". For highly sensitive questions ($W > .7$), a large value of T is needed to make this happen. As expected, $Var(\widehat{W})$, is always smaller for a one-stage model as compared to a two-stage model. Since the focus is more on effective estimation of μ_x, we recommend using a two-stage optional RRT model with small to moderate values of T for less sensitive questions and using higher values of T for more sensitive questions.

Table 5.7: Comparison of $V(\widehat{\mu}_x)$ (shown in bold) and $\widehat{V}(\widehat{\mu}_x)$ (not in bold) at various levels of W and T using optimum choices of n_1 and n_2 that minimize $\{V(\widehat{\mu}_x) + V(\widehat{W})\}$ for $\mu_x = 4$, $\sigma_x^2 = 4$, $\theta_1 = 2$, $\sigma_{S1}^2 = 2$, $\theta_2 = 5$, $\sigma_{S2}^2 = 5$

n	W	T = 0	T = .1	T = .3	T = .5	T = .7	T = .9
100	0.1	**0.28059**	**0.27514**	**0.26479**	**0.25447**	**0.24950**	**0.25665**
		0.27909	0.27429	0.25858	0.25720	0.25984	0.26553
	0.3	**0.36946**	**0.35893**	**0.33631**	**0.31088**	**0.28922**	**0.27538**
		0.37025	0.36049	0.34433	0.31994	0.29918	0.28027
	0.5	**0.41852**	**0.40995**	**0.38681**	**0.35660**	**0.32132**	**0.29058**
		0.38996	0.41574	0.40040	0.36909	0.33317	0.29728
	0.7	**0.42945**	**0.42988**	**0.41917**	**0.39161**	**0.35341**	**0.30574**
		0.41703	0.39911	0.43281	0.40764	0.36907	0.31006
	0.9	**0.40073**	**0.41852**	**0.43222**	**0.41559**	**0.37645**	**0.31639**
		0.39366	0.40787	0.41713	0.43285	0.38918	0.32196
500	0.1	**0.05610**	**0.05503**	**0.05289**	**0.05094**	**0.04998**	**0.05160**
		0.05509	0.05416	0.05174	0.05107	0.04788	0.05118
	0.3	**0.07392**	**0.07182**	**0.06722**	**0.06224**	**0.05775**	**0.05478**
		0.07720	0.07495	0.06842	0.05987	0.05587	0.05365
	0.5	**0.08370**	**0.08206**	**0.07741**	**0.07132**	**0.06448**	**0.05780**
		0.08352	0.08155	0.07798	0.07205	0.06216	0.05597
	0.7	**0.08582**	**0.08601**	**0.08378**	**0.07824**	**0.07043**	**0.06081**
		0.08312	0.08425	0.07941	0.07801	0.07113	0.05793
	0.9	**0.08009**	**0.08374**	**0.08633**	**0.08320**	**0.07555**	**0.06345**
		0.07899	0.08420	0.08498	0.08105	0.07711	0.05930
1000	0.1	**0.02805**	**0.02751**	**0.02643**	**0.02547**	**0.02497**	**0.02580**
		0.02811	0.02755	0.02708	0.02586	0.02534	0.02610
	0.3	**0.03695**	**0.03592**	**0.03362**	**0.03113**	**0.02885**	**0.02739**
		0.03614	0.03553	0.03316	0.03136	0.02948	0.02754
	0.5	**0.04185**	**0.04103**	**0.03872**	**0.03564**	**0.03224**	**0.02894**
		0.04026	0.04055	0.03777	0.03468	0.03230	0.02897
	0.7	**0.04290**	**0.04301**	**0.04189**	**0.03914**	**0.03519**	**0.03036**
		0.04131	0.04228	0.04188	0.03799	0.03401	0.03033
	0.9	**0.04005**	**0.04187**	**0.04316**	**0.04162**	**0.03774**	**0.03177**
		0.03726	0.03950	0.04204	0.04142	0.03713	0.03071

Table 5.8: Comparison of $V(\widehat{W})$ (shown in bold) and $\widehat{V}(\widehat{W})$ (not in bold) at various levels of W and T using optimum choices of n_1 and n_2 that minimize $\{V(\widehat{\mu}_x) + V(\widehat{W})\}$ for $\mu_x = 4$, $\sigma_x^2 = 4$, $\theta_1 = 2$, $\sigma_{S1}^2 = 2$, $\theta_2 = 5$, $\sigma_{S2}^2 = 5$

n	W	T = 0	T = .1	T = .3	T = .5	T = .7	T = .9
100	0.1	**0.02922**	**0.03498**	**0.05324**	**0.09558**	**0.23426**	**1.86223**
		0.02842	0.03389	0.05037	0.09484	0.24178	1.91306
	0.3	**0.04118**	**0.04889**	**0.07237**	**0.12382**	**0.28168**	**2.01198**
		0.04086	0.04743	0.07522	0.12753	0.28524	2.02801
	0.5	**0.04725**	**0.05699**	**0.08604**	**0.14539**	**0.32573**	**2.15743**
		0.04410	0.05777	0.08995	0.15113	0.33994	2.21298
	0.7	**0.04769**	**0.05957**	**0.09321**	**0.16196**	**0.36001**	**2.29585**
		0.04779	0.05736	0.09604	0.16995	0.38306	2.27753
	0.9	**0.04315**	**0.05674**	**0.09557**	**0.17448**	**0.39408**	**2.43178**
		0.04369	0.05726	0.09536	0.18236	0.40528	2.41371
500	0.1	**0.00586**	**0.00700**	**0.01071**	**0.01907**	**0.04677**	**0.37215**
		0.00590	0.00702	0.01060	0.01974	0.04578	0.36529
	0.3	**0.00821**	**0.00975**	**0.01452**	**0.02470**	**0.05643**	**0.40267**
		0.00886	0.01054	0.01521	0.02459	0.05452	0.39207
	0.5	**0.00945**	**0.01133**	**0.01716**	**0.02908**	**0.06492**	**0.43178**
		0.00911	0.01135	0.01774	0.03009	0.06357	0.41281
	0.7	**0.00960**	**0.01188**	**0.01869**	**0.03247**	**0.07224**	**0.45947**
		0.00926	0.01134	0.01797	0.03275	0.07434	0.43202
	0.9	**0.00869**	**0.01131**	**0.01922**	**0.03481**	**0.07855**	**0.48617**
		0.00847	0.01111	0.01850	0.03381	0.08142	0.44777
1000	0.1	**0.00293**	**0.00350**	**0.00537**	**0.00953**	**0.02341**	**0.18607**
		0.00279	0.00334	0.00526	0.00947	0.02354	0.18114
	0.3	**0.00411**	**0.00487**	**0.00725**	**0.01233**	**0.02824**	**0.20134**
		0.00399	0.00473	0.00692	0.01210	0.02774	0.19468
	0.5	**0.00473**	**0.00566**	**0.00857**	**0.01456**	**0.03246**	**0.21585**
		0.00452	0.00549	0.00820	0.01386	0.03114	0.20769
	0.7	**0.00481**	**0.00593**	**0.00935**	**0.01622**	**0.03615**	**0.22978**
		0.00467	0.00576	0.00918	0.01537	0.03326	0.21666
	0.9	**0.00434**	**0.00566**	**0.00961**	**0.01738**	**0.03931**	**0.24304**
		0.00411	0.00539	0.00935	0.01691	0.03715	0.21989

Table 5.9: Comparison of $V(\widehat{\mu}_x)$ (shown in bold) and $\widehat{V}(\widehat{\mu}_x)$ (not in bold) at various levels of W and T using equal sample sizes for $\mu_x = 4$, $\sigma_x^2 = 4$, $\theta_1 = 2$, $\sigma_{S1}^2 = 2$, $\theta_2 = 5$, $\sigma_{S2}^2 = 5$

n	W	T = 0	T = .1	T = .3	T = .5	T = .7	T = .9
100	0.1	**0.31333**	**0.30818**	**0.29760**	**0.28667**	**0.27538**	**0.26373**
		0.31203	0.30998	0.29768	0.29232	0.28027	0.26937
	0.3	**0.39778**	**0.38738**	**0.36418**	**0.33778**	**0.30818**	**0.27538**
		0.39283	0.38299	0.36641	0.33942	0.30998	0.28027
	0.5	**0.44667**	**0.43778**	**0.41333**	**0.38000**	**0.33778**	**0.28667**
		0.43499	0.43458	0.40860	0.38058	0.33942	0.29232
	0.7	**0.46000**	**0.45938**	**0.44507**	**0.41333**	**0.36418**	**0.29760**
		0.45704	0.45449	0.44233	0.40860	0.36641	0.29768
	0.9	**0.43778**	**0.45218**	**0.45938**	**0.43778**	**0.38738**	**0.30818**
		0.44233	0.44900	0.45449	0.43458	0.38299	0.30998
500	0.1	**0.06267**	**0.06164**	**0.05952**	**0.05733**	**0.05508**	**0.05275**
		0.06182	0.06035	0.05867	0.05604	0.05387	0.05158
	0.3	**0.07956**	**0.07748**	**0.07284**	**0.06756**	**0.06164**	**0.05508**
		0.08060	0.07915	0.07511	0.06626	0.06035	0.05387
	0.5	**0.08933**	**0.08756**	**0.08267**	**0.07600**	**0.06756**	**0.05733**
		0.09532	0.08618	0.08332	0.07707	0.06626	0.05604
	0.7	**0.09200**	**0.09188**	**0.08901**	**0.08267**	**0.07284**	**0.05952**
		0.09054	0.09201	0.08647	0.08332	0.07511	0.05867
	0.9	**0.08756**	**0.09044**	**0.09188**	**0.08756**	**0.07748**	**0.06164**
		0.09041	0.09309	0.09201	0.08618	0.07915	0.06035
1000	0.1	**0.03133**	**0.03082**	**0.02976**	**0.02867**	**0.02754**	**0.02637**
		0.03115	0.03077	0.03022	0.02911	0.02786	0.02661
	0.3	**0.03978**	**0.03874**	**0.03642**	**0.03378**	**0.03082**	**0.02754**
		0.03723	0.03693	0.03457	0.03281	0.03077	0.02786
	0.5	**0.04467**	**0.04378**	**0.04133**	**0.03800**	**0.03378**	**0.02867**
		0.04488	0.04193	0.03912	0.03608	0.03281	0.02911
	0.7	**0.04600**	**0.04594**	**0.04451**	**0.04133**	**0.03642**	**0.02976**
		0.04613	0.04573	0.04325	0.03912	0.03457	0.03022
	0.9	**0.04378**	**0.04522**	**0.04594**	**0.04378**	**0.03874**	**0.03082**
		0.04196	0.04509	0.04573	0.04193	0.03693	0.03077

Table 5.10: Comparison of $V(\widehat{W})$ (shown in bold) and $\widehat{V}(\widehat{W})$ (not in bold) at various levels of W and T using equal sample sizes for $\mu_x = 4$, $\sigma_x^2 = 4$, $\theta_1 = 2$, $\sigma_{S1}^2 = 2$, $\theta_2 = 5$, $\sigma_{S2}^2 = 5$

n	W	T = 0	T = .1	T = .3	T = .5	T = .7	T = .9
100	0.1	**0.02513**	**0.03019**	**0.04707**	**0.08647**	**0.22355**	**1.85713**
		0.02478	0.02989	0.04641	0.08856	0.22533	1.89456
	0.3	**0.03598**	**0.04281**	**0.06477**	**0.11331**	**0.27173**	**2.01198**
		0.03571	0.04181	0.06633	0.11320	0.26902	2.02801
	0.5	**0.04167**	**0.05028**	**0.07731**	**0.13500**	**0.31475**	**2.16167**
		0.04055	0.05012	0.07691	0.13606	0.31443	2.21403
	0.7	**0.04220**	**0.05259**	**0.08470**	**0.15153**	**0.35262**	**2.30620**
		0.04212	0.05248	0.08440	0.15075	0.36114	2.27419
	0.9	**0.03758**	**0.04975**	**0.08694**	**0.16291**	**0.38533**	**2.44558**
		0.03822	0.04995	0.08675	0.16238	0.37629	2.42121
500	0.1	**0.00503**	**0.00604**	**0.00941**	**0.01729**	**0.04471**	**0.37143**
		0.00490	0.00583	0.00909	0.01665	0.04332	0.35945
	0.3	**0.00720**	**0.00856**	**0.01295**	**0.02266**	**0.05435**	**0.40240**
		0.00729	0.00878	0.01354	0.02244	0.05247	0.38988
	0.5	**0.00833**	**0.01006**	**0.01546**	**0.02700**	**0.06295**	**0.43233**
		0.00854	0.00971	0.01543	0.02748	0.06232	0.41617
	0.7	**0.00844**	**0.01052**	**0.01694**	**0.03031**	**0.07052**	**0.46124**
		0.00816	0.01011	0.01625	0.03025	0.07373	0.44539
	0.9	**0.00752**	**0.00995**	**0.01739**	**0.03258**	**0.07707**	**0.48912**
		0.00767	0.00994	0.01671	0.03147	0.07905	0.47224
1000	0.1	**0.00251**	**0.00302**	**0.00471**	**0.00865**	**0.02236**	**0.18571**
		0.00234	0.00283	0.00451	0.00844	0.02168	0.17901
	0.3	**0.00360**	**0.00428**	**0.00648**	**0.01133**	**0.02717**	**0.20120**
		0.00321	0.00382	0.00565	0.01023	0.02544	0.19509
	0.5	**0.00417**	**0.00503**	**0.00773**	**0.01350**	**0.03148**	**0.21617**
		0.00424	0.00456	0.00693	0.01186	0.02840	0.21100
	0.7	**0.00422**	**0.00526**	**0.00847**	**0.01515**	**0.03526**	**0.23062**
		0.00428	0.00521	0.00784	0.01359	0.03077	0.22084
	0.9	**0.00376**	**0.00497**	**0.00869**	**0.01629**	**0.03853**	**0.24456**
		0.00368	0.00508	0.00862	0.01478	0.03442	0.22900

5.3 Concluding Remarks

This thesis attempts to carry forward the work in the area of RRT models, particularly in the area of sensitivity estimation. One can see that using the two independent sample technique works very well when it comes to simultaneous estimation of mean and sensitivity levels. Also, the additive RRT models provide more efficient estimators as compared to multiplicative RRT models. One should also note that the additive RRT models are more user-friendly. Another important issue is that of maintaining anonymity. There is a problem in using a multiplicative model since a non-zero reported response would generally indicate the presence of the sensitive behavior to some degree, thus revealing a respondent's true status in a face-to-face situation with the interviewer (therefore encouraging response distortion or non-response on the part of the respondent.) We also note that a two-stage optional RRT model generally produces lower $Var\left(\widehat{\mu}_x\right)$ but a somewhat higher $Var(\widehat{W})$ as compared to the one-stage model. However, as we remarked earlier, there is generally a greater emphasis on the estimation of μ_x. Hence a two-stage optional RRT model will be more effective than a one-stage optimal RRT model.

5.4 SAS Code

```
************************************************************;
* SAS code -- Two-Stage Additive Optional Model *
************************************************************;
%macro runit(par1,par2,par3,par4,par5,par6,par7);
data calc_ns;
 seed=&par1; theta_1=&par2; theta_2=&par3; mu_X=&par4; T=&par5;
 n=&par6; W=&par7;

/** sigma_sq_x = mu_x since poisson **/
/** sigma_sq_s1 = theta_1 since poisson **/
/** sigma_sq_s2 = theta_2 since poisson **/
sigma_sq_z1 = mu_X + theta_1*W*(1-T)
              + theta_1*theta_1*W*(1-T)*(1-W*(1-T));
sigma_sq_z2 = mu_X + theta_2*W*(1-T)
              + theta_2*theta_2*W*(1-T)*(1-W*(1-T));

/* calculate optimum n1 and n2 (that minimize variance) */
/* interim build vars */
sqrt_th1 = sqrt((theta_1*theta_1) + (1/((1-T)*(1-T))));
sqrt_th2 = sqrt((theta_2*theta_2) + (1/((1-T)*(1-T))));

n1 = round((n*(sqrt(sigma_sq_z1))*sqrt_th2)/
(((sqrt(sigma_sq_z2))*sqrt_th1)+((sqrt(sigma_sq_z1))*sqrt_th2)),1);
n2 = round((n*(sqrt(sigma_sq_z2))*sqrt_th1)/
(((sqrt(sigma_sq_z2))*sqrt_th1)+((sqrt(sigma_sq_z1))*sqrt_th2)),1);
```

```
chk_N = n1 + n2;

/* check */
*proc append base=chk_all data=calc_ns;run;
*proc print data = chk_all; *run;

data scramble_it (drop = z1bar z2bar) one;
 retain z1sum z1bar z2sum z2bar;
 set calc_ns (keep = seed T mu_X W theta_1 theta_2
                     sigma_sq_z1 sigma_sq_z2 n1 n2 n);

/* calculate proportion that would give true response */
TrueCombined = T + ((1-T)*(1-W));

do i=1 to 1000;
 /* Sample 1 */
 z1sum = 0; z1bar = 0;
 do j=1 to n1;
  * Poisson random variable X with mean mu_X;
   X = ranpoi(seed,mu_X);
  * Poisson random variable S1 with mean theta_1;
   S1 = ranpoi(seed,theta_1);

  /* Simulate True Responders with BERN Var using TrueCombined */
  TrueFlag = 0;
```

```
 if TrueCombined = 0 then TrueFlag=0;
                     else TrueFlag = ranbin(seed,1,TrueCombined);
 if TrueFlag = 1 then y1 = X;
                 else y1 = X + S1;
 Z1sum = Z1sum + y1;
output scramble_it;
end;
Z1bar = Z1sum/n1;     * average n1 response E(Z1);
j = .; S1 = .; y1 =.;   /* clean up */

/* Sample 2 */
z2sum = 0; z2bar = 0;
do k=1 to n2;
 X = ranpoi(seed,mu_X);
 S2 = ranpoi(seed,theta_2);

 TrueFlag = 0;
 if TrueCombined = 0  then TrueFlag=0;
                      else TrueFlag = ranbin(seed,1,TrueCombined);
 if TrueFlag = 1 then y2 = X;
                 else y2 = X + S2;
 Z2sum = Z2sum + y2;
output scramble_it;
end;
Z2bar = Z2sum/n2;     * average n2 response E(Z2);
k = .; S2 = .; y2 =.;   /* clean up */
```

```
output one;
end;
run;
/*check*/*proc means data = scramble_it mean std var; *run;

data final; set one; drop j k TrueFlag S1 S2 y1 y2 X /*b*/;
mu_X_hat = (z1bar*theta_2 - z2bar*theta_1)/(theta_2 - theta_1);
mu_W_hat = (z2bar - z1bar)/((theta_2 - theta_1)*(1-T));

/* calculate expected expected var(mu_X) and var(W) */
var_mu_X_calculated = (1/((theta_2 - theta_1)*(theta_2 - theta_1)))*
                       (((theta_2*theta_2)*(sigma_sq_z1/n1))
                       +((theta_1*theta_1)*(sigma_sq_z2/n2)));

var_mu_W_calculated = (1/((theta_1 - theta_2)*(theta_1 - theta_2)*
                      (1 - T)*(1 - T)))*((sigma_sq_z1/n1)
                      +(sigma_sq_z2/n2));

/* check */*proc means data = final mean std var;
*var  mu_X mu_X_hat W mu_W_hat; *run;

/* output variables */
proc means data = final mean std var noprint;
var    t    mu_X    n1    n2    n    W    mu_W_hat    mu_X_hat
       var_mu_X_calculated var_mu_W_calculated;
```

```
output out=Sumdata
 mean(t)=T        mean(mu_X)=TrueMuX    mean(W)=TrueW
 mean(n1)=n1      mean(n2)=n2           mean(n)=N
 mean(mu_W_hat)=SimulatedW              mean(mu_X_hat)=SimulatedMuX
 var(mu_X_hat)=SimulatedVarMuX
 mean(var_mu_X_calculated)=ExpectedVarMuX
 var(mu_W_hat)=SimulatedVarW
 mean(var_mu_W_calculated)=ExpectedVarW;
run;

data final_summary; set sumdata (drop=_type_);
proc append base=final_output data=final_summary;
proc print data = final_output; run;

%mend runit;

%runit(23,2,5,4,0,100,.1)
%runit(23,2,5,4,0,100,.2)
%runit(23,2,5,4,.1,100,.1)
%runit(23,2,5,4,.1,100,.2)
%runit(23,2,5,4,.3,100,.1)
%runit(23,2,5,4,.3,100,.2)
```

BIBLIOGRAPHY

[1] Crowne, D. P., and Marlowe, D. (1960), "A New Scale of Social Desirability Independent of Psychopathy", *Journal of Consulting Psychology*, **24**, 349-354.

[2] Eichhorn, B. H. and Hayre, L. S. (1983), "Scrambled Randomized Response Methods for Obtaining Sensitive Quantitative Data", *Journal of Statistical Planning and Inference*, **7**, 307-316.

[3] Greenberg, B. G., Abul-Ela, A. L. A., Simmons, W. R. and Horvitz, D. G. (1969), "The Unrelated Question Randomized Response Model - Theoretical Framework", *Journal of the American Statistical Association*, **64**, 520-539.

[4] Greenberg, B. G., Keubler, R. T., Jr., Abernathy, J. R., and Horvitz, D. G. (1971), "Application of Randomized Response Technique in Obtaining Quantitative Data", *Journal of the American Statistical Association*, **66**, 243-250.

[5] Gupta, S. N. (2001), "Quantifying the Sensitivity Level of Binary Response Personal Interview Survey Questions", *Journal of Combinatorics, Information and System Sciences*, **26**, 101-109.

[6] Gupta, S. N., Gupta, B. C. and Singh, S. (2002), "Estimation of Sensitivity Level of Personal Interview Survey Questions", *Journal of Statistical Planning and Inference*, **100**, 239-247.

[7] Gupta, S. N., and Thornton, B. (2002), "Circumventing Social Desirability Response Bias in Personal Interview Surveys", *American Journal of Mathematical and Management Sciences*, **22**, 369-383.

[8] Gupta, S. N., and Shabbir, J. (2004), "Sensitivity Estimation for Personal Interview Survey Questions", *Statistica*, **64**, 643-653.

[9] Gupta, S. N., Thornton, B., Shabbir, J., and Singhal, S. (2006), "A Comparison of Multiplicative and Additive Optional RRT Models", *Journal of Statistical Theory and Applications*, **5**, 226-239.

[10] Gupta, S. N. and Shabbir,J. (2007), "On the Estimation of Population Mean and Sensitivity in a Two-stage Optional Randomized Response Model", *Journal of Indian Society of Agricultural Statistics,* **61**, 164-168.

[11] Jones, E. E., and Sigall, H. (1971), "The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude", *Psychological Bulletin*, **76**, 349-364.

[12] Mangat, N. S. and Singh, R. (1990), "An Alternative Randomized Response Procedure", *Biometrika*, **77**, 439-442.

[13] Randall, D. M. and Fernandes, M. F. (1991), "The Social Desirability Response Bias in Ethics Research", *Journal of Business Ethics*, **10**, 805-817.

[14] Reynolds, W. M. (1982), "Development of a Reliable and Valid Short Form of the Marlowe-Crowne SDB Scale", *Journal of Clinical Psychology*, **38**, 119-125.

[15] Roese, N. J., and Jamieson, D. W. (1993), "Twenty Years of Bogus Pipeline Research: A Critical Review and Meta-analysis", *Psychological Bulletin*, **114**, 363-375.

[16] Ryu, J. B., Kim, J. M., Heo, T. Y., and Park, C. G. (2006), "On Stratified Randomized Response Sampling", *Model Assisted Statistics and Applications*, **1**, 31-36.

[17] Singhal, S. (2004), "Circumventing social desirability response bias and gender bias in personal interview surveys", Master's Thesis, University of Southern Maine.

[18] Thornton, B. and Gupta, S. N. (2004), "Comparative Validation of a Partial (versus Full) Randomized Response Technique: Attempting to Control for Social Desirability Response Bias to Sensitive Questions", *Individual Differences Research*, **2**, 214-224.

[19] Warner, S. L. (1965), "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", *Journal of the American Statistical Association*, **60**, 63-69.

[20] Warner, S. L. (1971), "The Linear Randomized Response Model", *Journal of the American Statistical Association*, **66**, 884-888.