PEREZ-SUAREZ, DAVID, M.A. Variations of the Greenberg Unrelated Question Binary Model. (2018)
Directed by Dr. Sat Gupta. 54 pp.

We explore different variations of the Greenberg Unrelated Question RRT model for a binary response (yes or no). In one variation, we allow m independent responses from each respondent. In another variation, we use inverse sampling where we record the number of responses leading up to the kth "yes" response. It turns out that for $m > 1$, the variance (theoretical and empirical) of the multiple independent response model decreases significantly relative to the regular Greenberg et al. (1969) model ($m = 1$). Furthermore, it was also noticed that for $k > 1$, the variance (theoretical and empirical) of the inverse sampling model decreases significantly as well relative to the inverse sampling model for $k = 1$. Thus, it turns out that both variations produce more efficient models. These results have been validated by theoretical comparisons, extensive computer simulations, and a field survey.

VARIATIONS OF THE GREENBERG UNRELATED QUESTION BINARY MODEL

by

David Perez-Suarez

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
2018

Approved by

_____
Committee Chair

APPROVAL PAGE

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

Committee Chair _____
Sat Gupta

Committee Members _____
Igor Erovenko

_____
Shan Suthaharan

_____
Haimeng Zhang

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

I would like to offer my most sincere gratitude to Dr. Sat Gupta for guiding me in every step of this project. I would also like to thank my committee members Professors Haimeng Zhang, Shan Suthaharan and Igor Erovenko for their valuable suggestions. Furthermore, I would like to thank all of the UNC-Greensboro Mathematics/Statistics professors for letting us in their classrooms so that we can carry out the fieldwork required for my project in an efficient manner. Lastly, I would like to thank my fellow graduate students Padma Manthena and Emily Johnson for helping me carry out the fieldwork. Without them, it would have taken much longer to carry it out.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER I

INTRODUCTION

1.1 Social Desirability Bias

Social desirability response bias (SDB) refers to the tendency of research subjects to give socially desirable responses instead of responses that represent their true feelings [7]. The bias in responses due to this personality trait becomes a major issue when the scope of the survey involves sensitive topics such as politics, religion, and environment; or personal issues such as drug use, cheating, and smoking [7]. This is evident throughout the literature.

Hebert et al. (1995) found that self-report of dietary intake could be biased by social desirability or social approval which in turn affected the risk estimates in epidemiological studies [10]. These constructs produced response set biases, which were seen when testing in domains characterized by easily desirable responses. Given the social and psychological value attributed to diet, assessment methodologies used most commonly in epidemiological studies are specifically vulnerable to these biases [10].

Fernandes and Randall (1992) found that social desirability was present when a questionnaire was administered under varying conditions of anonymity and with different measurement techniques for the social desirability construct [3]. Results showed that a social desirability bias was seen in the majority of relationships studied, and for the most part, played a little role [3]. While the measure of social desirability

affected the strength and relationship type, the condition of anonymity had relatively little effect on the level of social desirability [3].

Whenever possible, it is desirable to measure the extent of this bias present in responses to a survey by using a social desirability scale within the survey [7]. A number of methods to take care of this issue are suggested in the literature. One method that could help circumvent SDB is called the randomized response technique (RRT), which was introduced originally by Warner (1965) [15] and then generalized by other researchers such as Greenberg et al. (1969, 1971) [5] [6], Warner (1971) [16], Klein and Spady (1993) [11], Gupta et al. (2002) [8], and Gupta et al. (2010, 2013) [13] [9]. RRT Models allow sensitive information to be collected without showing the individual's sensitive status [15].

There are many types of RRT Models that are useful in interpreting survey data. However, for the sake of simplicity, we will concentrate on two types: regular RRT Models and optional RRT Models.

1.2 Regular RRT Models

The following sections present four models that are useful in understanding how regular RRT Models work: both the Warner's Binary Model (1965) and Warner's Quantitative Model (1971), and both the Greenberg et al. (1969) Binary Model and Greenberg et al. (1971) Quantitative Model.

1.2.1. Warner's Binary Model (1965)

Suppose we are interested in estimating the proportion of people who submitted an incorrect tax return last year on purpose. For this survey, we have a deck of cards that contains two types of questions:

Question 1. Did you submit an incorrect tax return last year?

Question 2. Did you submit a correct tax return last year?

A respondent picks Question 1 with probability p and picks Question 2 with probability $1 - p$, where p is known to the researcher. Thus, the respondent will say "yes" in two settings:

- when he/she submitted an incorrect tax return last year and the card picked contained Question 1
- when he/she submitted a correct tax return last year and the card picked contained Question 2

The observed proportion of "yes" responses in the sample is linearly related to the unknown proportion π of those who submitted an incorrect tax return last year on purpose as shown below.

Let p_y be the probability of a "yes" response [15]. Note that

$$p_y = p\pi + (1 - p)(1 - \pi).$$

We obtain the estimated value of p_y as $\hat{p_y} = \frac{n_1}{n}$ where n_1 is the number of respondents who answered "yes" from a sample size of n. Thus, our estimate of π ends up being the following.

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

The variance of this estimator, as given in Warner (1965) [15], is

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

The term $\frac{p(1-p)}{n(2p-1)^2}$ represents the penalty for using the randomized response model to estimate π. Furthermore, notice that this penalty is inversely proportional to n. Thus, choosing a larger n value leads to reducing the penalty.

1.2.2. Greenberg et al. (1969) Binary Model

As shown above, the initial Warner design involved the use of two related questions which divided the sample into two mutually exclusive and complementary sections. There is another randomized response design known as an unrelated question design which was introduced by Greenberg et al. (1969) [5].

In the unrelated-question model, the sample is divided into two mutually exclusive groups. It was suggested that the respondents' confidence in the anonymity provided by the technique might be further enhanced, and hence the veracity of their responses, if one of the two choices offered to the respondent referred to a non-sensitive, innocuous attribute, say Y, which was unrelated to the sensitive attribute A [5]. Two such binary questions (in statement form) might be:

- "Did you have an induced abortion last year?"
- "Were you born in the month of April?"

The respondent picks either of the two questions above based on the outcome of a random experiment. However, only one of the questions is related to the sensitive characteristic while the other question is not, due to the fact that the true proportion

of "yes" answers to the unrelated question would be either known or can be estimated from an independent survey.

In these models, there is no distinction as to whether an individual considers the associated question sensitive. Furthermore, the responses can be unscrambled at the aggregate level but not at the individual level [9]. Thus, unlike Warner's technique from the previous section, at least some of the respondents would have the reassurance that they answered a wholly unrelated question, resulting in more respondent cooperation than Warner's technique [9]. More detail (the theoretical aspect) for this method will be explained in Chapter III.

1.2.3. Warner's Quantitative Model (1971)

Warner (1971) proposed several quantitative RRT models including the following model.

Let X represent the true sensitive variable of interest with unknown mean μ_X and unknown variance σ_X^2, and S represent a scrambling variable with known true mean μ_S and known variance σ_S^2, where S is independent of X [16]. Let Y represent the reported response. Thus

$$Y = X + S.$$

The expected response is then given by

$$\begin{aligned} E(Y) &= E(X) + E(S) \\ &= \mu_X + \mu_S. \end{aligned}$$

Estimating $E(Y)$ by the sample mean $(\bar{Y})$ of the reported responses, an unbiased estimator for the mean of the sensitive variable is obtained as shown below

$$\hat{\mu_X} = \bar{Y} - \mu_S.$$

The variance of this estimator is given by

$$\begin{aligned} Var(\hat{\mu}_X) &= Var(\bar{Y}) \\ &= \frac{\sigma_Y^2}{n} \\ &= \frac{\sigma_X^2}{n} + \frac{\sigma_S^2}{n}. \end{aligned}$$

In the above equation, the $\frac{\sigma_S^2}{n}$ term is the penalty for using the RRT model.

1.2.4. Greenberg et al. (1971) Quantitative Model

Greenberg et al. (1971) Quantitative Model is a quantitative version of Greenberg et al. (1969) Binary Model.

In this model, a known proportion of respondents (p) is given the sensitive question with response A randomly, and the remaining proportion $(1-p)$ of the respondents is given a non-sensitive question with response B where the mean (μ_B) of the non-sensitive variable is known [6].

Let a sample of size n be drawn with replacement. Also, let the reported response be Z [6]. Then we have

$$Z = \begin{cases} A & \text{with probability p} \\ B & \text{with probability 1-p, and} \end{cases}$$

The expected value of Z is given by

$$E(Z) = p\mu_A + (1-p)\mu_B.$$

Using the expression for $E(Z)$ from above, μ_A can be estimated using the sample mean of reported responses ($\bar{Z}$) [6]. Estimating $E(Z)$ by $\bar{Z}$, we obtain

$$\hat{\mu}_A = \frac{\bar{Z} - (1-p)\mu_B}{p}.$$

The variance of the estimator above is given by

$$Var(\hat{\mu}_A) = \frac{\sigma_Z^2}{np^2},$$

where $\sigma_Z^2 = \sigma_Y^2 + \sigma_S^2[(1-T)W] + \sigma^2[(1-T)W]\{1-[(1-T)W]\}$.

1.3 Optional RRT Models

For this section, we discuss three models that will help us in understanding how optional RRT Models work: Gupta et al. (2002) [8] Multiplicative Optional

Model, Gupta et al. (2010) [13] Two-Stage Additive Optional Model and Gupta et al. (2013) [9] Optional Unrelated-Question Randomized Response Model.

1.3.1. Gupta et al. (2002) [8] Multiplicative Optional Model

Gupta et al. (2002) [8] came up with multiplicative optional scrambling in which a simple random sample of size n with replacement was taken and within this sample, an unknown proportion (W) of respondents scramble their responses in the case of perceiving the question as sensitive, and reporting the actual response otherwise.

Let Z denote the reported response and Y denote the true response of the sensitive study variable. Suppose that there was a deck of cards that was provided to the respondent. Each card within the deck has numbers that follow a known probability distribution, S, which is independent of Y, which has mean $E(S) = 1$ and known variance σ_S^2. The respondent picks a card and gives a scrambled response in the case he/she sees the question as sensitive, and reports only Y otherwise. Let W denote the probability of reporting a multiplicatively scrambled response where W is an unknown parameter and is known as the sensitivity level of the research question. Therefore, in this model, there are two unknown parameters that need to be estimated: μ_Y and W. Also,

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

which leads to

$$E(Z) = E(Y).$$

Taking into account the sample mean of the reported responses ($\bar{Z}$), we can obtain $\hat{\mu}_Y$ (an estimate of μ_Y) which is given by:

$$\hat{\mu}_Y = \bar{Z}.$$

Furthermore, notice that

$$\begin{aligned} Var(\hat{\mu}_Y) &= \frac{1}{n} Var(Z) \\ &= \frac{1}{n}[\sigma_Y^2 + W\sigma_S^2(\sigma_Y^2 + \mu_Y^2)]. \end{aligned}$$

Also, note that

$$Z = S^T Y \text{ where } T \sim Bernoulli(W).$$

Using this relation, the parameter W is estimated as follows:

$$\begin{aligned} ln(Z) &= Tln(S) + ln(Y) \\ E(ln(Z)) &= E(T)E(ln(S)) + E(ln(Y)) \end{aligned}$$

Using the first order Taylor's approximation of $E(ln(Y))$, we obtain

$$E(ln(Z)) \approx WE(ln(S)) + ln(E(Y))$$
$$W \approx \frac{E(ln(Z)) - ln(E(Y))}{E(ln(S))}$$

Thus, an estimator of W is given by:

$$\hat{W} = \frac{\frac{1}{n}\sum_{i=1}^{n} ln(Z_i) - ln(\frac{1}{n}\sum_{i=1}^{n} Y_i)}{E(ln(S))}.$$

1.3.2. Gupta et al. (2010) [13] Two-Stage Additive Optional Model

Let Y denote a quantitative sensitive variable with unknown mean μ_Y and unknown variance σ_Y^2, where μ_Y is to be estimated [13]. Let the sample size n be split into two sub-samples of sizes n_1 and n_2 where $n_1 + n_2 = n$. Let S_i be the scrambling variable used to scramble the responses in the ith sub-sample ($i = 1, 2$) where it is assumed that Y and S_i are mutually independent. Let θ_i denote the known mean for S_i and $\sigma_{S_i}^2$ is the known variance. Within each sub-sample, a predetermined proportion (T) of respondents are directed to give out the true response to the question being asked and the remaining proportion ($1-T$) of respondents give an additively scrambled response if they think the question is sensitive, and a true response otherwise. Let W denote the sensitivity level of the underlying sensitive question. For this model, we obtain the reported response in the ith sub-sample (Z_i)

as

$$Z_i = \begin{cases} Y & \text{with probability T+(1-W)(1-T)} \\ Y + S_i & \text{with probability W(1-T)}, \quad i = 1, 2. \end{cases}$$

The mean for Z_i $(i = 1, 2)$ is given by [8]:

$$E(Z_i) = \mu_Y + \theta_i W(1 - T), \ \textit{where } E(S_i) = \theta_i \ \ (i = 1, 2).$$

It follows from above that

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

The above expressions lead to the following unbiased estimators of μ_Y and W:

$$\hat{\mu}_Y = \frac{\theta_2 \bar{Z}_1 - \theta_1 \bar{Z}_1}{\theta_2 - \theta_1}, \theta_1 \neq \theta_2,$$
$$\hat{W} = \frac{\bar{Z}_2 - \bar{Z}_1}{(\theta_2 - \theta_1)(1 - T)}, \theta_1 \neq \theta_2, T \neq 1,$$

where $\bar{Z}_i$ $(i = 1, 2)$ represents the sample mean of responses in the ith $(i = 1, 2)$ sub-sample [8].

Note that

$$Var(\hat{\mu}_Y) = \frac{1}{(\theta_2 - \theta_1)^2}[\frac{\theta_2^2}{n_1}\sigma_{Z_1}^2 + \frac{\theta_1^2}{n_2}\sigma_{Z_2}^2],$$

and

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

where

$$\sigma_{Z_i}^2 = \sigma_Y^2 + \sigma_{S_i}^2[(1-T)W] + \sigma_i^2[(1-T)W]\{1 - [(1-T)W]\}.$$

1.3.3. Gupta et al. (2013) [9] Optional Unrelated-Question Randomized Response Models

Gupta et al. (2013) [9] proposed a generalization of the Unrelated Question RRT techniques. We take into account both the binary and quantitative response situations and estimate the prevalence (π) of the sensitive behavior and the mean response (μ) of the quantitative sensitive question. In addition, the model also estimates the sensitivity level (W) of the underlying question, which is the proportion of subjects who consider the question to be sensitive, and hence choose to give a scrambled response. The following provides the theoretical background of the quantitative and binary models, respectively.

We proceed first with the quantitative version of this model. Let X represent the true sensitive variable of interest with unknown mean μ_X and unknown variance σ_X^2, and Y represent the a non-sensitive variable with known mean μ_Y and known variance σ_Y^2. Furthermore, let p represent the probability of getting the sensitive question from the randomization device. Let W represent the sensitivity level of the question as before. Thus, the reported response Z is given by the following.

$$Z = \begin{cases} X & \text{with probability } (1-W) + Wp \\ Y & \text{with probability } W(1-p) \end{cases}$$

with

$$E(Z) = (1-W)E(X) + W(pE(X) + (1-p)E(Y)), \text{ and}$$
$$Var(Z) = [(1-W) + Wp]E(X^2) + W(1-p)E(Y^2) - [E(Z)]^2$$

Thus, to solve the above equation for the two unknown parameters μ_X and W, we use a split-sample approach where the total sample size is divided into two sub-samples. Each sub-sample uses a randomization device with a different probability $(p_i, i = 1, 2)$ of getting the sensitive question. The expected response in the ith $(i = 1, 2)$ sub-sample then is given by:

$$E(Z_i) = (1-W)E(X) + W(p_iE(X) + (1-p_i)E(Y)), \textit{ where i=1,2.}$$

Solving the two equations above, we get the following:

$$\frac{E(Z_1) - E(X)}{E(Z_2) - E(X)} = \frac{1 - p_1}{1 - p_2}.$$

Solving for $E(X)$, we get

$$E(X) = \frac{E(Z_1) - \lambda E(Z_2)}{1 - \lambda}, \text{ where } \lambda = \frac{1 - p_1}{1 - p_2}.$$

This leads to estimating μ_X by

$$\hat{\mu}_X = \frac{\bar{Z}_1 - \bar{Z}_2}{1 - \lambda}.$$

The variance of this estimator is given by

$$\begin{aligned} var(\mu_X) &= \frac{var(\bar{Z}_1 + \lambda \bar{Z}_2)}{(1 - \lambda)^2}, \\ &= \frac{\frac{Var(Z_1)}{n_1} + \lambda^2 \frac{Var(Z_2)}{n_2}}{(1 - \lambda)^2}, \end{aligned}$$

where

$$var(Z_i) = \frac{[(1 - W) + W p_i] E(X^2) + W(1 - p_i) E(Y^2) - [E(Z_i)]^2}{n_i}, \ i = 1, 2.$$

We also estimate the proportion of subjects who scramble their responses (W). From the formula for $E(Z_i)$, we get the following for the estimator of W.

$$\hat{W} = \frac{\bar{Z}_1 - \bar{Z}_2}{\mu_Y(p_2 - p_1) + (1 - p_2)\bar{Z}_1 - (1 - p_1)\bar{Z}_2},$$

Using first-order Taylor's approximation where $A = E(\bar{Z}_1)$ and $B = E(\bar{Z}_2)$, we get the following.

$$\begin{aligned}
\hat{W} &= \hat{W}(A, B) + \left.\frac{\partial \hat{W}(Z_1, Z_2)}{\partial \bar{Z}_1}\right|_{A,B} (\bar{Z}_1 - A) + \left.\frac{\partial \hat{W}(\bar{Z}_1, \bar{Z}_2)}{\partial \bar{Z}_2}\right|_{A,B} (\bar{Z}_2 - B) \\
&= \frac{A - B}{\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B} + \frac{(p_2 - p_1)(\mu_Y - B)(\bar{Z}_1 - A)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]^2} \\
&\quad + \frac{(p_2 - p_1)(A - \mu_Y)(\bar{Z}_2 - B)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]^2} \\
&= \hat{W}_1
\end{aligned}$$

We obtain the following for $\mathrm{E}(\hat{W}_1)$.

$$\begin{aligned}
E(\hat{W}_1) &= \frac{A - B}{\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B} \\
&\quad + \frac{(p_2 - p_1)(\mu_Y - B)(E(\bar{Z}_1) - A)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]^2} \\
&\quad + \frac{(p_2 - p_1)(A - \mu_Y)(E(\bar{Z}_2) - B)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]^2} \\
&= \frac{A - B}{\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B} \\
&= W
\end{aligned}$$

The associated variance of $\hat{W}_1$ is given by

$$var(\hat{W}_1) = (\frac{(p_2 - p_1)(\mu_Y - B)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]^2})^2 \frac{\sigma_1^2}{n_1} + (\frac{(p_2 - p_1)(\mu_Y - B)}{[\mu_Y(p_2 - p_1) + (1 - p_2)A - (1 - p_1)B]^2})^2 \frac{\sigma_2^2}{n_2},$$

where $\sigma_i^2 = [1 - W + Wp_i]E(X^2) + W(1 - p_i)E(Y^2) - [E(Z_i)]^2, i = 1, 2.$

We now proceed with the binary version for this model. In many cases the main research interest is in the sensitive attribute. In this case, the research question requires a binary response where the answer could be either "yes" or "no".

Let X represent a sensitive binary variable of interest with unknown mean π_X, and Y represent a non-sensitive binary variable with known mean π_Y. Let p represent the probability of receiving the sensitive question from the randomization device. Thus, the probability of a "yes" response (p_Y) is given by:

$$p_Y = (1 - W)\pi_Y + W[p\pi_Y + (1 - p)\pi_Y]$$

Like before, the sample is divided into two sub-samples to solve for π_X and W. In this case, the probability of a "yes" response in the ith ($i = 1, 2$) sub-sample is given by

$$p_{Y\,i} = (1 - W)\pi_Y + W[p_i\pi_Y + (1 - p_i)\pi_Y], \quad i = 1, 2.$$

From the above equation, we get the following.

$$\pi_X = \frac{p_{Y1} - \lambda p_{Y2}}{1-\lambda} \text{ where } \lambda = \frac{1-p_1}{1-p_2}$$

From the equation of π_X, we get the following for the estimator of π_X.

$$\hat{\pi}_X = \frac{\hat{p}_{Y_1} - \lambda \hat{p}_{Y_2}}{1-\lambda},$$

with the variance of this estimator given by

$$Var(\hat{\pi}_X) = \frac{Var(\hat{p}_{Y_1}) + \lambda^2 Var(\hat{p}_{Y_2})}{(1-\lambda)^2},$$

where $Var(\hat{p}_{Y_i}) = \frac{p_{Y_i}(1-Y_i)}{n_i} \quad (i = 1, 2)$.

From the equation for π_x, we find an estimator for W in the binary case as

$$\hat{W}_\pi = \frac{\hat{p}_{Y_1} - \hat{p}_{Y_2}}{\pi_Y(p_2 - p_1) + (1-p_2)\hat{p}_{Y_1} - (1-p_1)\hat{p}_{Y_2}}$$

Applying the first order Taylors approximation expansion for a bivariate function where $A = E(\hat{p}_{Y_1})$ and $B = E(\hat{p}_{Y_2})$, this can be approximated as follows:

$$\begin{aligned}\hat{W}_\pi \approx & \frac{A-B}{\pi_Y(p_2-p_1) + (1-p_2)A - (1-p_1)B} + \frac{(p_2-p_1)(\pi_Y - B)(\hat{p}_{Y_1} - A)}{[\pi_Y(p_2-p_1) + (1-p_2)A - (1-p_1)B]^2} \\ & + \frac{(p_2-p_1)(A-\pi_Y)(\hat{p}_{Y_2} - B)}{[\pi_Y(p_2-p_1) + (1-p_2)A - (1-p_1)B]^2} \\ = & \hat{W}_1(\text{say})\end{aligned}$$

Thus, the mean for $\hat{W}_1$ is the following.

$$\begin{aligned} E(\hat{W}_1) &= \frac{A-B}{\pi_Y(p_2-p_1)+(1-p_2)A-(1-p_1)B} \\ &= W_\pi \end{aligned}$$

Its associated variance is given by

$$\begin{aligned} var(\hat{W}_1) = &(\frac{(p_2-p_1)(\pi_Y-B)}{[\pi_Y(p_2-p_1)+(1-p_2)A-(1-p_1)B]^2})^2\frac{\sigma_1^2}{n_1} \\ &+(\frac{(p_2-p_1)(\pi_Y-B)}{[\pi_Y(p_2-p_1)+(1-p_2)A-(1-p_1)B]^2})^2\frac{\sigma_2^2}{n_2} \end{aligned}$$

where $\sigma_i^2 = \frac{p_{Y_i}(1-p_{Y_i})}{n_i}, i = 1, 2$

CHAPTER II

APPLICATIONS OF RRT MODELS

2.1 RRT Applications

As it was mentioned before, Warner (1965) introduced the RRT models. RRT models allow sensitive information to be collected without showing any single individual's sensitive status by the use of a randomization device that leads to more accurate estimates of sensitive behaviors [9]. RRT models have been used quite a lot in field surveys as shown below.

2.1.1. Abernathy et al. (1970)-Estimates of Induced Abortion Rates in North Carolina

According to the American Congress of Obstetricians and Gynecologists (AGOG) website, an induced abortion is when an action is done or medication is taken to end a pregnancy. Consequences of induced abortions include increased risk of premature birth, increased risk of stillbirth, heavy bleeding and severe pain. Abernathy et al. (1970) [1] used RRT models to obtain estimates of induced abortion rates within urban North Carolina. In this case, estimates of the proportion of women aged 18-44 years that had an induced abortion during the past year were reported [1]. For the study, population indices were made relating induced abortion with total conceptions in terms of whites and nonwhites [1].

It was found that the illegal abortion rate per 100 conceptions was approximately 14.9 for whites and 32.9 for nonwhites [1]. Furthermore, as mentioned before, estimates of the proportion of women aged 18 years or older that had an abortion during their lifetime are also shown [1]. Among married women, the proportion hav-

ing an abortion during their lifetime declined as education increased where estimates were high for women with 5 or more pregnancies [1]. As a whole, most of the respondents stated that they were satisfied that the randomized response approach would not reveal their personal situation [1]. Looking closer, it was evident that woman respondents would have hesitated to respond truthfully to the sensitive question of induced abortions [1].

2.1.2. Striegel et al. (2006)-Doping and Drug Use in Elite Sports

According to the UNESCO website, doping refers to an athlete's use of prohibited drugs or methods to improve training and sporting results. As a person ventures in a sport, doping is a very important issue that an athlete will face due to their being tested for drugs, the continuing use of drugs and some of their competitors potentially cheating due to the continued use of drugs. All athletes who were questioned were subjected to doping controls as members or junior members of the national teams [14].

In order to estimate the prevalence of doping and illicit drug abuse, the athletes were either issued an anonymous standardized questionnaire (SQ) or were interviewed using randomized response technique (RRT) where there were 1394 participants in the SQ group and 480 participants in the RRT group [14]. Official doping tests showed 0.81% of athletes testing positive for doping, while 6.8% of the athletes confessed to having practiced doping based on the RRT model [14]. In addition and more importantly, the present RRT study revealed an alarmingly high prevalence of illicit drug use, specifically of cocaine use, that has been severely underestimated by previous studies.[14].

2.1.3. Ostapczuck et al. (2009)-The Education Effect in Attitudes towards Foreigners

Even though a negative correlation has been found between a respondent's education and his negative attitude towards foreigners, the reasons for this education effect are still unknown and up for debate [12]. In order to study this relationship in Germany, Ostapczuck et al. (2009) postulated a hypothesis that stated the highly educated may not be genuinely less xenophobic, but more prone to give socially desirable, xenophile answers in attitude questionnaires [12].

In order to examine this hypothesis, two survey methods were used: direct questioning and RRT [12]. Under direct questioning, 75% of the highly educated expressed a xenophile attitude, whereas only 55% of the less educated expressed the same type of attitude [12]. Under the RRT part, we obtained significantly reduced estimates of 53% for the proportion of xenophiles who were highly educated, and 24% among those who were less educated, showing a strong distortion of self-reported attitudes towards foreigners within both groups [12]. However, when we look at the results closer, these two survey methods demonstrated great variation [12].

2.1.4. Gill et al. (2013)-Estimates of Risky Sexual Behaviors

Gupta et al. (2013) introduced optional unrelated question RRT models for both the binary and quantitative response to surveys that carried sensitive questions. Both the mean estimator of the sensitive variable and the prevalence estimator of the sensitive characteristic carried the property of asymptotic normality. Also, in each case, the sensitivity level estimator, using first order Taylor's approximation, carried the property of asymptotic normality as well. These mathematical results were verified through the use of computer simulations. The binary and quantitative response models were used in surveys that had a sensitive behavior attached to check

that these results were true in fieldwork applications [4]. The binary question of interest was "Have you been told by a healthcare professional that you have a sexually transmitted disease?", whereas the quantitative question of interest was "How many sexual partners have you had in the last 12 months?" [4].

The survey was conducted using three methods: optional unrelated question RRT method, direct face-to-face interviewing and anonymous check-box survey method [4].

Table 1. Estimates of the Mean Number of Sexual Partners in the Last 12 Months

Method	$\hat{\mu}$	sample std. dev.	95% CI*	n
Optional RRT	1.717	3.9912	(1.2744, 2.1596)	466
Check-Box	1.680	2.5613	(1.2647, 2.0953)	218
Face-to-Face Interview	1.130	1.1511	(0.9311, 1.3289)	192

* Based on Bonferroni Correction

Table 2. Estimates of the STD Diagnosis Prevalence

Method	$\hat{\pi_X}$	sample std. dev.	95% CI*	n
Optional RRT	0.0367	0.1180	(0.0159, 0.0576)	466
Check-Box	0.0900	0.2862	(0.0438, 0.1362)	220
Face-to-Face Interview	0.0200	0.1400	(-0.0042, 0.0442)	192

* Based on Bonferroni Correction

Based upon Tables (1) and (2) given in [4], it was observed that the optional unrelated question RRT method's estimates were closer to the anonymous check-box survey method's estimates, and the lowest point estimate was obtained by face-to-face interview method, which is expected as it provided the lowest anonymity [4].

2.1.5. Chhabra et al. (2016)-Estimating Prevalence of Sexual Abuse by an Acquaintance

More recently, Chhabra et al. (2016) [2] used RRT models to estimate the prevalence of sexual abuse of college female students by either a friend or an acquaintance. Looking closely, Chhabra et al. (2016) [2] used RRT models to simultaneously estimate the mean of a sensitive variable and the sensitivity level of the main sensitive question without the use of the traditional split-sample approach. The data were collected from a survey that was conducted by the authors on a sample of undergraduate female students aged 17-21 years at a college of University of Delhi, India from January 2015 [2].

For this study, the binary research question of interest (Question 1) was "Have you ever been a victim of sexual abuse by friend or family member?" and the quantitative research question (Question 2) was "How many days in a typical month do you watch pornographic clips/videos/movies on movie channels, WhatsApp, Youtube, Internet etc.?" [2]. It was found that within the confidential survey, 27 out of 195 subjects answered "yes" [2]. Also, within the face-to-face survey, 16 out of 195 subjects replied reassuringly, and within the optional unrelated question RRT method, 40 out of 195 subjects responded positively for Question 1 whereas 25 out of 195 subjects responded positively for Question 2 [2]. These led to the following estimates of π and μ.

Table 3. Estimates of Sexual Abuse Prevalence

Method	$\hat{\pi}$	$\hat{Var}(\hat{\pi})$
Confidential	0.138461	0.00061174
Face-to-Face Survey	0.082051	0.00038600
Optional RRT Model	0.119021	0.00067577

Table 4. Estimates of Mean Number of Days the Respondent Watch Porn Clips

Method	$\hat{\mu_X}$	$\hat{Var}(\hat{\mu_X})$
Confidential	3.216	0.825404
Face to Face Survey	1.560	1.026010
Optional RRT Model	2.765	0.011500

CHAPTER III

VARIATIONS OF GREENBERG ET AL. (1969) MODEL

Our focus is on the Greenberg et al. (1969) Binary Model. In the first variation, we allow a respondent to provide multiple independent responses and in the second variation, we use a technique called inverse sampling. From these two variations, we derive two estimators and compare them with the regular Greenberg et al. (1969) Model. The second estimator involved inverse sampling while waiting for the first "yes" response, and also inverse sampling while waiting for the kth "yes" response.

3.1 Greenberg et al. (1969) Model with Multiple Independent Responses

Let us first recall the Greenberg et al. (1969) Unrelated Question RRT Model but in more detail. Let π_x be the unknown prevalence of a sensitive attribute X in the population and π_y be the known prevalence of a non-sensitive attribute Y. A randomization device offers respondents a choice between two questions, the sensitive question and an unrelated question with respective probabilities p and $1-p$. Let p_y be the probability of a "yes" response. Then

$$p_y = \pi_x p + \pi_y(1-p), \tag{3.1}$$

which leads to the estimator

$$\hat{\pi}_G = \frac{\hat{p}_y - \pi_y(1-p)}{p}, \tag{3.2}$$

where $\hat{p}_y$ is the sample proportion of "yes" responses.

The mean of the estimator in (3.2) is given by

$$E(\hat{\pi}_G) = \pi_x, \tag{3.3}$$

which signifies that $\hat{\pi}_G$ is an unbiased estimator of π_x.

The variance of the estimator in (3.2) is given by

$$Var(\hat{\pi}_G) = \frac{p_y(1-p_y)}{np^2}. \tag{3.4}$$

Now, suppose each respondent is allowed to give m independent responses in a SR-SWR (simple random sample with replacement) of size n. Let T_i be the number of "yes" responses provided by the ith respondent. Then

$$T_i \sim \text{Binomial(m, } p_y) \text{ where } \; p_y = \pi_x p + \pi_y(1-p), \text{ and } E(T_i) = mp_y.$$

If $\bar{T} = \frac{\sum T_i}{n}$, then we know that $E(\bar{T}) = mp_y$.

Estimating mp_y by $\bar{T}$, the estimator for π_x in (3.2) can be refined as below:

$$\hat{\pi}_{GM} = \frac{\frac{\bar{T}}{m} - (1-p)\pi_y}{p}. \tag{3.5}$$

Note that

$$\begin{aligned} E(\hat{\pi}_{GM}) &= \frac{\frac{E(\bar{T})}{m} - (1-p)\pi_y}{p} \\ &= \frac{\frac{mp_y}{m} - (1-p)\pi_y}{p} \\ &= \frac{p_y - (1-p)\pi_y}{p} \\ &= \pi_x. \end{aligned} \tag{3.6}$$

The variance of the estimator $\hat{\pi}_{GM}$ is given by

$$\begin{aligned} Var(\hat{\pi}_{GM}) &= \frac{1}{m^2p^2} Var(\bar{T}) \\ &= (\frac{1}{m^2p^2})(\frac{mp_y(1-p_y)}{n}) \\ &= \frac{p_y(1-p_y)}{nmp^2}. \end{aligned} \tag{3.7}$$

3.2 Inverse Sampling-Waiting for First "Yes" Response

Suppose we continue to use the Greenberg et al. (1969) Model until a "yes" response is recorded. Let S_i be the total number of trials needed in the ith run to

reach the first "yes" response. Then,

$$S_i \sim \text{Geometric } (p_y), \tag{3.8}$$

with $E(S_i) = \frac{1}{p_y}$ and $Var(S_i) = \frac{1-p_y}{p_y^2}$, where p_y is defined in (3.1).

Also, let there be a SRSWR of n respondents and $\bar{S}$ be the sample mean of S_i's. Then $\frac{1}{p_y}$ can be estimated by $\bar{S}$ which leads to $\hat{p}_y = \frac{1}{\bar{S}}$ being an estimator of p_y. Using first order Taylor's approximation of $\frac{1}{\bar{S}}$, we obtain

$$\frac{1}{\bar{S}} \approx \frac{1}{E(S)} + (\bar{S} - E(S))(\frac{-1}{(E(S))^2}) \tag{3.9}$$

where $E(S) = \frac{1}{p_y}$.

With this approximation,

$$\begin{aligned} E(\frac{1}{\bar{S}}) &\approx \frac{1}{E(S)} \\ &= p_y \end{aligned} \tag{3.10}$$

Then, using $\frac{1}{\bar{S}}$ as an estimator of p_y, the estimator in (3.2) becomes

$$\hat{\pi}_{GI} = \frac{\frac{1}{\bar{S}} - (1-p)\pi_y}{p} \tag{3.11}$$

Note that

$$\begin{aligned} E(\hat{\pi}_{GI}) &= \frac{E(\frac{1}{S}) - (1-p)\pi_y}{p} \\ &\approx \frac{p_y - (1-p)\pi_y}{p} \\ &= \pi_x. \end{aligned} \tag{3.12}$$

since $E(\frac{1}{S}) \approx p_y$, as argued in (3.10).

Thus, we see that $\hat{\pi}_{GI}$ is an unbiased estimator of π_x, up to first order of approximation.

From (3.11),

$$Var(\hat{\pi}_{GI}) = \frac{1}{p^2} Var(\frac{1}{\bar{S}}) \tag{3.13}$$

But

$$
\begin{aligned}
Var(\frac{1}{\bar{S}}) &\approx Var(\frac{1}{E(\bar{S})} + (\bar{S} - E(\bar{S}))(-\frac{1}{(E(\bar{S}))^2})) \\
&= Var(-\bar{S}p_y^2) \quad \text{from (3.10)} \\
&= p_y^4 Var(\bar{S}) \\
&= p_y^4 \frac{Var(S)}{n} \\
&= p_y^4 (\frac{\frac{1-p_y}{p_y}}{np_y}) \\
&= p_y^4 (\frac{1-p_y}{np_y^2}) \\
&= \frac{p_y^2(1-p_y)}{n}. \qquad (3.14)
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
Var(\hat{\pi}_{GI}) &\approx \frac{1}{p^2}(\frac{p_y^2(1-p_y)}{n}) \\
&= \frac{p_y^2(1-p_y)}{np^2}. \qquad (3.15)
\end{aligned}
$$

3.3 Inverse Sampling-Waiting for Kth "Yes" Response

Let $S_{i,k}$ be the total number of trials needed to reach the kth "yes" response in the ith run. Then, we see that $S_{i,k} \sim$ Negative Binomial(p_y,k) with

$$
E(S_{i,k}) = \frac{k}{p_y} \qquad (3.16)
$$

and

$$Var(S_{i,k}) = \frac{k(1-p_y)}{p_y^2}, \tag{3.17}$$

Also, let there be a sample of n independent trials on S, and $\bar{S}$ be the sample mean of the n $S_{i,k}$ values. Then

$$\begin{aligned} E(\bar{S}) &= E(S_{i,k}) \\ &= \frac{k}{p_y}. \end{aligned} \tag{3.18}$$

Therefore, $\frac{k}{p_y}$ can be estimated by $\bar{S}$ and $\hat{p_y} = \frac{k}{\bar{S}}$ can be used as an estimator of p_y. Using first-order Taylor's approximation,

$$\frac{1}{\bar{S}} = \frac{1}{E(S)} + (\bar{S} - E(S))(-\frac{1}{(E(S))^2}), \tag{3.19}$$

where $E(S) = \frac{k}{p_y}$

Thus, our estimator for π_x in (3.2) becomes:

$$\hat{\pi}_{GI_k} = \frac{\frac{k}{\bar{S}} - \pi_y(1-p)}{p}. \tag{3.20}$$

From (3.20), we get the following for the mean of $\hat{\pi}_{GI_k}$.

$$\begin{aligned} E(\hat{\pi}_{GI_k}) &= \frac{kE(\frac{1}{S}) - \pi_y(1-p)}{p} \\ &\approx \frac{k(\frac{p_y}{k}) - \pi_y(1-p)}{p} \\ &= \frac{p_y - \pi_y(1-p)}{p} \\ &= \pi_x \end{aligned} \tag{3.21}$$

Thus, $\hat{\pi}_{GI_k}$ is an unbiased estimator of π_x, up to first order of approximation.

From (3.20), we also get,

$$Var(\hat{\pi}_{GI_k}) = \frac{k^2}{p^2} Var(\frac{1}{\bar{S}}) \tag{3.22}$$

But,

$$\begin{aligned} Var(\frac{1}{\bar{S}}) &\approx Var(\frac{1}{E(\bar{S})} + (\bar{S} - E(\bar{S}))(-\frac{1}{(E(\bar{S}))^2})) \\ &= \frac{p_y^4}{k^4} Var(\bar{S}) \quad \text{from (3.18)} \\ &\approx \frac{p_y^4}{k^4} (\frac{\frac{k(1-p_y)}{p_y^2}}{n}) \\ &= \frac{p_y^4}{k^4} (\frac{k(1-p_y)}{np_y^2}) \\ &= \frac{p_y^2(1-p_y)}{k^3 n}, \end{aligned} \tag{3.23}$$

Thus, we have:

$$\begin{aligned} Var(\hat{\pi}_{GI_k}) &\approx \frac{k^2}{p^2}(\frac{p_y^2(1-p_y)}{k^3 n}) \\ &= \frac{p_y^2(1-p_y)}{knp^2} \end{aligned} \tag{3.24}$$

CHAPTER IV

EFFICIENCY COMPARISONS

In this chapter, we compare the efficiencies of the following estimators that were discussed in the previous chapter:

$\hat{\pi}_G$ = Greenberg et al. (1969) estimator

$\hat{\pi}_{GM}$ = Greenberg et al. (1969) estimator using m independent responses

$\hat{\pi}_{GI}$ = Greenberg et al. (1969) estimator using inverse sampling waiting for the first "yes" response

$\hat{\pi}_{GI_k}$ = Greenberg et al. (1969) estimator using inverse sampling waiting for the kth "yes" response

4.1 Greenberg et al. (1969) Model vs. Greenberg et al. (1969) Multiple Response Model

We can summarize (4.1) as follows.

$$\begin{aligned} Var(\hat{\pi}_{GM}) &= \frac{p_y(1-p_y)}{nmp^2} \\ &= \frac{1}{m}(\frac{p_y(1-p_y)}{np^2}) \\ &= \frac{1}{m}Var(\hat{\pi}_G) \end{aligned} \tag{4.1}$$

Based on (4.1), note that $Var(\hat{\pi}_{GM}) < Var(\hat{\pi}_G)$ for $m > 1$. Thus, the Greenberg et al. (1969) Multiple Response Model is more efficient than the regular Greenberg et al. (1969) Model.

4.2 Inverse Sampling Stopping at the First "Yes" Response Model vs. Greenberg et al. (1969) Model

We can summarize (4.2) as follows.

$$\begin{aligned} Var(\hat{\pi}_{GI}) &= \frac{p_y^2(1-p_y)}{np^2} \\ &= p_y(\frac{p_y(1-p_y)}{np^2}) \\ &= p_y Var(\hat{\pi}_G) \end{aligned} \tag{4.2}$$

Based on (4.2), note that $Var(\hat{\pi}_{GI}) < Var(\hat{\pi}_G)$ for $p_y < 1$. Thus, the Inverse Sampling-Waiting for the First "Yes" Response Model is always more efficient than the regular Greenberg et al. (1969) Model.

4.3 Inverse Sampling Stopping at the Kth "Yes" Response Model vs. Greenberg et al. (1969) Model

We can summarize (4.3) as follows.

$$\begin{aligned} Var(\hat{\pi}_{GI_k}) &= \frac{p_y^2(1-p_y)}{nkp^2} \\ &= p_y(\frac{p_y(1-p_y)}{nkp^2} \\ &= \frac{p_y}{k}(\frac{p_y(1-p_y)}{np^2}) \\ &= \frac{p_y}{k}Var(\hat{\pi}_G) \end{aligned} \tag{4.3}$$

Based on (4.3), note that $Var(\hat{\pi}_{GI_k}) < Var(\hat{\pi}_G)$ for $k > 1$ and $p_y < 1$. Thus, the Inverse Sampling-Waiting for the Kth "Yes" Response Model is more efficient than the regular Greenberg et al. (1969) Model when $k > 1$.

4.4 Inverse Sampling Stopping at the First "Yes" Response Model vs. Greenberg et al. (1969) Multiple Response Model

We can summarize (4.4) as follows.

$$\begin{aligned} Var(\hat{\pi}_{GI}) &= \frac{p_y^2(1-p_y)}{np^2} \\ &= p_y(\frac{p_y(1-p_y)}{np^2}) \\ &= mp_y(\frac{p_y(1-p_y)}{nmp^2}) \\ &= mp_yVar(\hat{\pi}_{GM}) \end{aligned} \tag{4.4}$$

Based on (4.4), note that $Var(\hat{\pi}_{GI}) < Var(\hat{\pi}_{GM})$ for $mp_y < 1$. Thus, the Inverse Sampling-Waiting for the First "Yes" Response Model is more efficient than the Greenberg et al. (1969) Multiple Response Model when $m < \frac{1}{p_y}$.

4.5 Inverse Sampling Stopping at the Kth "Yes" Response Model vs. Greenberg et al. (1969) Multiple Response Model

We can summarize (4.5) as follows.

$$\begin{aligned} Var(\hat{\pi}_{GI_k}) &= \frac{p_y^2(1-p_y)}{nkp^2} \\ &= \frac{p_y}{k}(\frac{p_y(1-p_y)}{np^2}) \\ &= \frac{mp_y}{k}(\frac{p_y(1-p-y)}{nmp^2}) \\ &= \frac{mp_y}{k}Var(\hat{\pi}_{GM}) \end{aligned} \tag{4.5}$$

Based on (4.5), note that $Var(\hat{\pi}_{GI_k}) < Var(\hat{\pi}_{GM})$ for $mp_y < 1$ and $k > 1$. Thus, the Inverse Sampling-Waiting for the Kth "Yes" Response Model is more efficient than the Greenberg et al. (1969) Multiple Response Model when $k > 1$ and $m < \frac{1}{p_y}$.

4.6 Inverse Sampling Stopping at the Kth "Yes" Response Model vs. Inverse Sampling Stopping at the First "Yes" Response Model

We can summarize (4.6) as follows.

$$\begin{aligned} Var(\hat{\pi}_{GI_k}) &= \frac{p_y^2(1-p_y)}{nkp^2} \\ &= \frac{1}{k}(\frac{p_y^2(1-p_y)}{np^2}) \\ &= \frac{1}{k}Var(\hat{\pi}_{GI}) \end{aligned} \tag{4.6}$$

Based on (4.6), note that $Var(\hat{\pi}_{GI_k}) < Var(\hat{\pi}_{GI})$ for $k > 1$. Thus, the Inverse Sampling-Waiting for K "Yes" Responses Model is more efficient than the Inverse Sampling-Waiting for the First "Yes" Response Model when $k > 1$.

4.7 Summary

We can summarize sections (4.1)-(4.6) as follows:

$$\begin{aligned} Var(\hat{\pi}_{GI_k}) &< Var(\hat{\pi}_{GI}) && \text{if } k > 1 \\ &< Var(\hat{\pi}_{GM}) && \text{if } mp_y < 1 \\ &< Var(\hat{\pi}_{G}) && \text{if } m > 1 \end{aligned} \tag{4.7}$$

Overall, based on (4.7), the Inverse Sampling-Waiting for K "Yes" Responses Model is more efficient than the Inverse Sampling-Waiting for the First "Yes" Response Model, the Greenberg et al. (1969) Model with Multiple Independent Responses Model and

the regular Greenberg et al. (1969) Model. Thus, the Inverse Sampling-Waiting for K "Yes" Responses Model is the most efficient out of all our models when $k > 1$.

CHAPTER V

SIMULATION RESULTS

5.1 Simulation Process

All of the theoretical formulas from Chapter 3 were tested empirically through the use of computer simulations to see how well the three estimators that were mentioned above compared against the regular Greenberg et al. (1969) Model. In order to carry out these simulations, we used different values of n, π_x, π_y p and a total of 10000 simulations and organized these results in 4 tables to see if the same conclusion was reached each time. Refer to the Results section of this chapter to see these tables.

5.2 Results

Table (5) below presents simulation results that were obtained from SAS for a total of 10000 simulations with sample sizes of 100, 500 and 1000, $\pi_x = 0.30$, $\pi_y = 0.7$ and $p = 0.85$.

Table (6) below presents simulation results that were obtained from SAS for a total of 10000 simulations with a sample size of 500, $\pi_x = 0.30$, $\pi_y = (0.5, 0.75, 0.9)$ and $p = 0.85$.

Table (7) below presents simulation results that were obtained from SAS for a total of 10000 simulations with a sample size of 500, $\pi_x = (0.40, 0.50, 0.70)$, $\pi_y = 0.7$ and $p = 0.85$.

Table (8) below presents simulation results that were obtained from SAS for a total of 10000 simulations with a sample size of 1000, $\pi_x = 0.50$, $\pi_y = 0.5$ and $p = 0.85$.

Table 5. Estimators of π_x with Corresponding Empirical ($\hat{Var}(\hat{\pi})$) and Theoretical ($Var(\hat{\pi})$) Variances with $n = (100, 500, 1000)$, $\pi_x = 0.30$, $\pi_y = 0.7$ and $p = 0.85$

n	m	$\hat{\pi}_{GM}$	$\hat{Var}(\hat{\pi}_{GM})$	$Var(\hat{\pi}_{GM})$
100	1	0.3003682	0.0031673	0.003188927
100	2	0.2998106	0.0015793	0.001594464
100	3	0.2996055	0.0010796	0.001062976
100	4	0.3003759	0.000802351	0.000797232
100	5	0.2998591	0.000652692	0.000637785
	k	$\hat{\pi}_{GI_k}$	$\hat{Var}(\hat{\pi}_{GI_k})$	$Var(\hat{\pi}_{GI_k})$
100	1	0.3027742	0.0011788	0.001148014
100	2	0.3013239	0.000575621	0.000574007
100	3	0.3007781	0.000384914	0.000382671
100	4	0.3007940	0.000284774	0.000287003
100	5	0.3005728	0.000229412	0.000229603
	m	$\hat{\pi}_{GM}$	$\hat{Var}(\hat{\pi}_{GM})$	$Var(\hat{\pi}_{GM})$
500	1	0.3001781	0.000638371	0.000637785
500	2	0.3002513	0.000318622	0.000318893
500	3	0.3000282	0.000214628	0.000212595
500	4	0.3000342	0.000156454	0.000159446
500	5	0.3000586	0.000126368	0.000127557
	k	$\hat{\pi}_{GI_k}$	$\hat{Var}(\hat{\pi}_{GI_k})$	$Var(\hat{\pi}_{GI_k})$
500	1	0.3004394	0.000229236	0.000229603
500	2	0.3002714	0.000112837	0.000114801
500	3	0.3000640	0.000076382	0.000076534
500	4	0.3000161	0.000058789	0.000057401
500	5	0.3002176	0.000046028	0.000045921
	m	$\hat{\pi}_{GM}$	$\hat{Var}(\hat{\pi}_{GM})$	$Var(\hat{\pi}_{GM})$
1000	1	0.3000929	0.000322274	0.000318893
1000	2	0.2998735	0.000159751	0.000159446
1000	3	0.2997484	0.000104373	0.000106298
1000	4	0.3000094	0.000079531	0.000079723
1000	5	0.2999271	0.000063045	0.000063779
	k	$\hat{\pi}_{GI_k}$	$\hat{Var}(\hat{\pi}_{GI_k})$	$Var(\hat{\pi}_{GI_k})$
1000	1	0.3001464	0.000117352	0.000114801
1000	2	0.3002190	0.000055895	0.000057401
1000	3	0.3000042	0.000038364	0.000038267
1000	4	0.3000767	0.000028260	0.000028700
1000	5	0.3000507	0.000022996	0.000022960

Table 6. Estimators of π_x with Corresponding Empirical ($\hat{Var}(\hat{\pi})$) and Theoretical ($Var(\hat{\pi})$) Variances with $n = 500$, $\pi_x = 0.30$, $\pi_y = (0.5, 0.75, 0.9)$ and $p = 0.85$

π_y	m	$\hat{\pi}_{GM}$	$\hat{Var}(\hat{\pi}_{GM})$	$Var(\hat{\pi}_{GM})$
0.5	1	0.3002275	0.000626069	0.000612042
0.5	2	0.3001521	0.000312660	0.000306021
0.5	3	0.3002485	0.000204204	0.000204014
0.5	4	0.2999008	0.000151733	0.000153010
0.5	5	0.2998362	0.000123436	0.000122408
	k	$\hat{\pi}_{GI_k}$	$\hat{Var}(\hat{\pi}_{GI_k})$	$Var(\hat{\pi}_{GI_k})$
0.5	1	0.3005469	0.000200526	0.000201974
0.5	2	0.3003312	0.000100022	0.000100987
0.5	3	0.3002731	0.000067926	0.000067325
0.5	4	0.3001613	0.000050848	0.000050493
0.5	5	0.3001627	0.000041351	0.000040395
	m	$\hat{\pi}_{GM}$	$\hat{Var}(\hat{\pi}_{GM})$	$Var(\hat{\pi}_{GM})$
0.75	1	0.3000440	0.000630890	0.000643443
0.75	2	0.2998798	0.000325955	0.000321721
0.75	3	0.3001845	0.000209802	0.000214481
0.75	4	0.2999747	0.000162359	0.000160861
0.75	5	0.2998438	0.000127722	0.000128689
	k	$\hat{\pi}_{GI_k}$	$\hat{Var}(\hat{\pi}_{GI_k})$	$Var(\hat{\pi}_{GI_k})$
0.75	1	0.3005070	0.000236190	0.000236465
0.75	2	0.3002626	0.000118541	0.000118233
0.75	3	0.3003537	0.000079194	0.000078822
0.75	4	0.3001062	0.000060759	0.000059116
0.75	5	0.3000874	0.000047707	0.000047293
	m	$\hat{\pi}_{GM}$	$\hat{Var}(\hat{\pi}_{GM})$	$Var(\hat{\pi}_{GM})$
0.9	1	0.2998793	0.000656826	0.000658547
0.9	2	0.3000140	0.000333343	0.000329273
0.9	3	0.2997773	0.000222884	0.000219516
0.9	4	0.3000899	0.000164766	0.000164637
0.9	5	0.2999541	0.000132506	0.000131709
	k	$\hat{\pi}_{GI_k}$	$\hat{Var}(\hat{\pi}_{GI_k})$	$Var(\hat{\pi}_{GI_k})$
0.9	1	0.3007569	0.000260245	0.000256833
0.9	2	0.3002132	0.000130448	0.000128417
0.9	3	0.3001158	0.000084538	0.000085611
0.9	4	0.3000875	0.000065457	0.000064208
0.9	5	0.3001639	0.000051279	0.000051367

Table 7. Estimators of π_x with Corresponding Empirical ($\hat{Var}(\hat{\pi})$) and Theoretical ($Var(\hat{\pi})$) Variances with $n = 500$, $\pi_x = (0.40, 0.50, 0.70)$, $\pi_y = 0.7$ and $p = 0.85$

π_x	m	$\hat{\pi}_{GM}$	$\hat{Var}(\hat{\pi}_{GM})$	$Var(\hat{\pi}_{GM})$
0.40	1	0.4001351	0.000673737	0.000683668
0.40	2	0.4002534	0.000339815	0.000341834
0.40	3	0.4000639	0.000226107	0.000227889
0.40	4	0.3999487	0.000174955	0.000170917
0.40	5	0.3999861	0.000138396	0.000136734
	k	$\hat{\pi}_{GI_k}$	$\hat{Var}(\hat{\pi}_{GI_k})$	$Var(\hat{\pi}_{GI_k})$
0.40	1	0.4006873	0.000297300	0.000304232
0.40	2	0.4003551	0.000152255	0.000152116
0.40	3	0.4001929	0.000098417	0.000101411
0.40	4	0.4001444	0.000076796	0.000076058
0.40	5	0.4001647	0.000060247	0.000060846
	m	$\hat{\pi}_{GM}$	$\hat{Var}(\hat{\pi}_{GM})$	$Var(\hat{\pi}_{GM})$
0.50	1	0.5001941	0.000689617	0.000689550
0.50	2	0.4999305	0.000342844	0.000344775
0.50	3	0.4998585	0.000240454	0.000229850
0.50	4	0.4999967	0.000175033	0.000172388
0.50	5	0.5000457	0.000139796	0.00013791
	k	$\hat{\pi}_{GI_k}$	$\hat{Var}(\hat{\pi}_{GI_k})$	$Var(\hat{\pi}_{GI_k})$
0.50	1	0.5007369	0.000364217	0.000365462
0.50	2	0.5002518	0.000180320	0.000182731
0.50	3	0.5004557	0.000121947	0.000121821
0.50	4	0.4999440	0.000089609	0.000091365
0.50	5	0.5000173	0.000072907	0.000073092
	m	$\hat{\pi}_{GM}$	$\hat{Var}(\hat{\pi}_{GM})$	$Var(\hat{\pi}_{GM})$
0.70	1	0.6995186	0.000571947	0.000581315
0.70	2	0.6999876	0.000285606	0.000290657
0.70	3	0.7001192	0.000196976	0.000193772
0.70	4	0.7000021	0.000145367	0.000145329
0.70	5	0.6995873	0.000119359	0.000116263
	k	$\hat{\pi}_{GI_k}$	$\hat{Var}(\hat{\pi}_{GI_k})$	$Var(\hat{\pi}_{GI_k})$
0.70	1	0.7006316	0.000397176	0.000406920
0.70	2	0.7002850	0.000202967	0.000203460
0.70	3	0.7001467	0.000132911	0.000135640
0.70	4	0.7003175	0.000102237	0.000101730
0.70	5	0.7000754	0.000080616	0.000081384

Table 8. Estimators of π_x with Corresponding Empirical ($Var(\hat{\pi})$) and Theoretical ($\hat{Var}(\hat{\pi})$) Variances with $n = 1000$, $\pi_x = 0.50$, $\pi_y = 0.5$ and $p = 0.85$

$\hat{\pi}_G$	$\hat{Var}(\hat{\pi}_G)$	$Var(\hat{\pi}_G)$	m	$\hat{\pi}_{GM}$	$\hat{Var}(\hat{\pi}_{GM})$	$Var(\hat{\pi}_{GM})$
0.4997153	0.000341679	0.000346021	1	0.4997153	0.000341679	0.000346021
			2	0.4998329	0.000174297	0.000173010
			3	0.4999238	0.000115451	0.000115340
			4	0.4998771	0.000085898	0.000086505
			5	0.4999516	0.000071178	0.000069204
$\hat{\pi}_{GI}$	$\hat{Var}(\hat{\pi}_{GI})$	$Var(\hat{\pi}_{GI})$	k	$\hat{\pi}_{GI_k}$	$\hat{Var}(\hat{\pi}_{GI_k})$	$Var(\hat{\pi}_{GI_k})$
0.5002282	0.000180216	0.000173010	1	0.5002282	0.000180216	0.000173010
			2	0.5000227	0.000085718	0.000086505
			3	0.5001590	0.000057470	0.000057670
			4	0.5000500	0.000044477	0.000043253
			5	0.5000404	0.000034781	0.000034602

Based on Tables (5)-(8), notice that as m increased, the variances of $\hat{\pi}_{GM}$ decreased relative to the variances of $\hat{\pi}_G$ which supports our formulas for the theoretical variance for both $\hat{\pi}_G$ and $\hat{\pi}_{GM}$ in chapter III. On the other hand, as k increased, the variances of $\hat{\pi}_{GI_k}$ decreased relative to the variances of $\hat{\pi}_{GI}$ which supports our formulas for the theoretical variance for both $\hat{\pi}_{GI}$ and $\hat{\pi}_{GI_k}$ in chapter III. Furthermore, looking at the results of $\hat{\pi}_{GI}$ relative to $\hat{\pi}_{GM}$, as mp_y increased, the variances of $\hat{\pi}_{GM}$ decreased relative to the variances of $\hat{\pi}_{GI}$, which supports our efficiency comparison of $\hat{\pi}_{GM}$ vs. $\hat{\pi}_{GI}$ from chapter IV.

5.3 Conclusion

Based on Tables (5)-(8), we can see that the regular Greenberg et al. (1969) model has a higher variance (theoretical and empirical) than the modified Greenberg model with multiple responses and those involving inverse sampling. Hence, the proposed variations of the Greenberg et al. (1969) model are more efficient than the

regular Greenberg et al. (1969) model. Looking closer at Tables (5)-(8), we can see that the modified Greenberg model with inverse sampling has a lower variance than that involving multiple independent responses. Hence, the modified Greenberg model with inverse sampling is the most efficient model out of all our models. However, greater effort is needed in using these newer models. Given that the gain in efficiency with newer models is quite substantial, the newer models are worth trying. However, in practice, we need to keep m and k small such as $m \leq 3$ and $k \leq 3$ since we want the respondent not to feel tired or exhausted of participating in the survey for a long time.

CHAPTER VI

FIELDWORK VALIDATION

In this chapter, we explore the fieldwork validation for our proposed variations of the Greenberg Unrelated Question RRT Model: the regular Greenberg et al. (1969) Model, the Greenberg Model with Multiple Independent Responses and the Inverse Sampling Model while Waiting for the First "Yes" Response, to check how they compare to the Anonymous group.

6.1 Procedure

In order to carry out this fieldwork, we had to first get approval from the IRB (Institutional Review Board). The approval process from the IRB took roughly an entire semester since we had to come up with the pertinent forms (student consent forms, student letter, etc.) and describe in careful detail what we wanted to accomplish in this fieldwork. Also, we had to go through the human subjects training to see if we had the necessary skills to do research on humans. Once we got approved and got the human subjects training done, we started the fieldwork process. For this fieldwork, we used a team of four people: Dr. Sat Gupta, Emily Johnson, Padma Manthera and myself. As a team, we recruited students from various classes with the permission from the instructors where the recruited students consisted of current UNC-Greensboro students from the mathematics and statistics classes offered at the time. The students were given a student letter that explained what the fieldwork entailed and an informed consent form of their approval to be in this fieldwork since participation was strictly voluntary. The recruited students were asked to partici-

pate in a survey associated with a medical condition. The survey question was if the student had ever been told by a healthcare professional that she/he has a sexually transmitted disease. Recruited students were randomly chosen to be in one of four groups: an Anonymous group, a group that corresponded to the original Greenberg RRT Model, a group that used the Greenberg Model based on 3 Independent Responses and a group that used the Greenberg Model based on Inverse Sampling while Waiting for the First "Yes" Response.

6.2 Groups Used

6.2.1. Group 1-Anonymous Group

The first group we used for the fieldwork was the Anonymous group. For this group, the respondent is given a piece of paper by the researcher, which contains the sensitive question. The respondent answers the question, and then the paper is put into a box.

6.2.2. Group 2-Greenberg et al. (1969) Model

The second group we used for this fieldwork was the regular Greenberg et al. (1969) Model. For this group, the researcher makes a deck of 100 cards where 85% of the deck contains the sensitive question and 15% of the deck contains an unrelated question ("Were you born in January-April?"). The researcher shuffled this deck and in turn, the respondent had to pick a card and answer the associated question. The researcher then records the answer to the question on a datasheet associated with Group 2.

6.2.3. Group 3-Greenberg et al. (1969) Model with 3 Independent Responses

The third group we used for this fieldwork was the Greenberg Model with Multiple Independent Responses. For this group, the researcher has the same deck as before. The researcher shuffles this deck three different times and in turn, the respondent has to pick three cards and give his/her responses. The researcher then records the responses to the three cards they picked on a datasheet associated with Group 3.

6.2.4. Group 4-Inverse Sampling while Waiting for First "Yes" Response

The fourth group we used for this fieldwork was the Inverse Sampling Model while Waiting for the First "Yes" Response. For this group, the researcher has the same deck as before. The researcher shuffles this deck and in turn, the respondent answers the card they picked. There were practical constraints in executing the survey under the conditions of Group 4 where we were to record the number of trials needed to get to the first "yes" response. This would have required a very large sample. So, we simply looked at results of Group 3 one more time and observed the number of times a "yes" response was reached. Then, we recorded the number of trials needed to get to each "yes" response.

6.3 Results

Once we obtained the data for the four groups, we put them in Excel spreadsheets separately. Then, we analyzed the data for each group by using the appropriate formulas from these four groups to see how they compare against each other. For these groups, the sample sizes varied depending on the student's response rate. Specifically,

the sample size for Group 4 depended on the responses from Group 2 since we had to count how many responses were there until the first "yes" response was reached.

Table 9. Estimated Prevalence (π) and Variance of $\hat{\pi}$ from the Four Groups

Estimator	n	$\hat{\pi}$	Variance
$\hat{\pi}_A$	241	0.04564315	0.0001807463
$\hat{\pi}_G$	178	0.04037607	0.0006000411
$\hat{\pi}_{GM}$	216	0.05742992	0.0001901209
$\hat{\pi}_{GI}$	15	0.04093619	0.0006065249

For Group 4, we looked at the results for Group 3 and observed there were 15 sequences of responses ending with a "yes" response.

According to the Odyssey website, 1 in 4 college students are infected with STD which means that the prevalence of STD among college students is 0.25. Comparing the prevalence of STD among the four groups with the above prevalence, we see that the prevalence of STD among the four groups (Greenberg et al. (1969), etc.) is much lower than the above prevalence for STD (0.25).

6.4 Conclusion

When we compare the point estimators (mean) from each group in Table (9), we notice some interesting results. When we compare the Anonymous group with the regular Greenberg et al. (1969) Model and the Inverse Sampling Model, we see that the Anonymous group had a higher prevalence of STD than those for the regular Greenberg et al. (1969) Model and the Inverse Sampling Model. However, when we compare the Anonymous group with the Greenberg Multiple Response Model, we see that the Anonymous group had a lower prevalence of STD than that for the Greenberg Multiple Response Model.

When we compare the variances from each group in Table (9), we notice some interesting results as well. Comparing the Anonymous group with the regular Greenberg et al. (1969) group, the Greenberg Multiple Response and the Inverse Sampling groups, the Anonymous group had a lower variance than that for these groups. Interestingly, the Multiple Response Model also proved very efficient. The Inverse Sampling Model would be the most efficient if we could ensure a comparable sample size.

CHAPTER VII

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

The main objective of this study was to explore variations of the famous Greenberg et al. (1969) Model. It turns out that all of the proposed variations are more efficient than the original model, with the model based on inverse sampling with $k > 1$ being the best. Admittingly, this is the most difficult to implement also.

Computer simulations validate over the mathematical findings. The fieldwork survey adds to this validation.

7.2 Future Work

The problem with the original unrelated-question RRT model is that it cannot differentiate on whether an individual actually considers the topic sensitive; every subject is assumed to find the research question sensitive, so all subjects utilize the randomization device to produce a scrambled response [9]. However, a topic or question may be sensitive for one person, but not sensitive for another. As it was mentioned before, optional RRT models, introduced by Gupta et al. [2002] [8], take this into account by allowing subjects who do not find the question sensitive to answer it without using the randomization step [9]. Subjects who find the research question sensitive still use the randomization device prior to giving a response [9]. In this model, the researcher does not know as to whether or not the subject used the scrambling device or gave a truthful response [9]. In the future, we hope to extend the variations that we came up with for the Greenberg et al. (1969) Model by introducing

optionality into them to see how well they compare against the regular Greenberg et al. (1969) Model.

REFERENCES

[1] James R Abernathy, Bernard G Greenberg, and Daniel G Horvitz. Estimates of induced abortion in urban north carolina. *Demography*, 7(1):19–29, 1970.

[2] Anu Chhabra, BK Dass, and Sat Gupta. Estimating prevalence of sexual abuse by an acquaintance with an optional unrelated question rrt model. *The North Carolina Journal of Mathematics and Statistics*, 2:1–9, 2016.

[3] Maria F Fernandes and Donna M Randall. The nature of social desirability response effects in ethics research. *Business Ethics Quarterly*, pages 183–205, 1992.

[4] Tracy Spears Gill, Anna Tuck, Sat Gupta, Mary Crowe, and Jennifer Figueroa. A field test of optional unrelated question randomized response models: estimates of risky sexual behaviors. In *Topics from the 8th Annual UNCG Regional Mathematics and Statistics Conference*, pages 135–146. Springer, 2013.

[5] Bernard G Greenberg, Abdel-Latif A Abul-Ela, Walt R Simmons, and Daniel G Horvitz. The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64(326):520–539, 1969.

[6] Bernard G Greenberg, Roy R Kuebler Jr, James R Abernathy, and Daniel G Horvitz. Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66(334):243–250, 1971.

[7] Pamela Grimm. Social desirability bias. *Wiley International Encyclopedia of Marketing*, 2010.

[8] Sat Gupta, Bhisham Gupta, and Sarjinder Singh. Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and inference*, 100(2):239–247, 2002.

[9] Sat Gupta, Anna Tuck, Tracy Gill, and Mary Crowe. Optional unrelated-question randomized response models. *Involve, a Journal of Mathematics*, 6(4):483–492, 2013.

[10] James R Hebert, Lynn Clemow, Lori Pbert, Ira S Ockene, and Judith K Ockene. Social desirability bias in dietary self-report may compromise the validity of

dietary intake measures. *International journal of epidemiology*, 24(2):389–398, 1995.

[11] Roger W Klein and Richard H Spady. An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421, 1993.

[12] Martin Ostapczuk, Jochen Musch, and Morten Moshagen. A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology*, 39(6):920–931, 2009.

[13] Javid Shabbir and Sat Gupta. Estimation of the finite population mean in two phase sampling when auxiliary variables are attributes. *Hacettepe Journal of Mathematics and Statistics*, 39(1), 2010.

[14] Heiko Striegel, Perikles Simon, Jochen Hansel, Andreas M Niess, and Rolf Ulrich. Doping and drug use in elite sports: An analysis using the randomized response technique: 1626: Board# 265 9: 30 am–10: 30 am. *Medicine & Science in Sports & Exercise*, 38(5):S247, 2006.

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

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