View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by The University of North Carolina at Greensboro

SIHM, JEONG SEP, Ph.D. Modified Binary Randomized Response Technique Models. (2017)
Directed by Dr. Sat Narain Gupta. 102 pp.

Social Desirability Bias (SDB) is the tendency in respondents to answer questions untruthfully in the hope of giving good impression to others. SDB occurs when the survey question is highly sensitive or personal, and responses cause sample statistics to systematically overestimate or underestimate corresponding population parameters. The Randomized Response Technique (RRT) is one of several methods to get around SDB in surveys involving sensitive questions in a face-to-face interview.

We first review some of the well-established binary response RRT models including the two-parameter models such as the two-stage RRT model and the optional RRT model. Then, we examine an optional RRT model based on the unrelated question RRT as presented by Gupta, Tuck, Spears Gill, and Crowe (2013). Also, we show another optional RRT model based on the two-stage RRT. Next, we carry out efficiency comparisons between these models and show simulation results. While these two models are all based on the split-sample approach to estimate two unknown parameters of interest ($\pi$ and $\omega$—the prevalence of sensitive characteristic and the sensitivity level of the underlying question respectively), the next two models utilize the two-question approach instead. One of them relies on the unrelated question RRT model. And the other relies on the two-stage optional RRT model. Again, efficiencies of estimators are compared and simulation results are provided.

In the end, simulation results and figures are presented and some conclusions are made regarding which estimator performs better. It turns out that the two-stage optional indirect RRT model with two-question approach performs better than other binary optional RRT models.

# MODIFIED BINARY RANDOMIZED RESPONSE TECHNIQUE MODELS

by

Jeong Sep Sihm

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

Greensboro
2017

Approved by

Committee Chair

*To my wife,*

*Mi-Young,*

*as well as to my three delightful children,*

*Jin Woo, Jane, and Jay Young,*

*for their support and encouragement.*

## APPROVAL PAGE

This dissertation written by Jeong Sep Sihm 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 Narain Gupta

Committee Members ______________________________
Gregory Copeland Bell

______________________________
Xiaoli Gao

______________________________
Lakshmi Sundaram Iyer

______________________________
Haimeng Zhang

______________________________
Date of Acceptance by Committee

______________________________
Date of Final Oral Examination

# ACKNOWLEDGMENTS

The work for this dissertation was undertaken at the suggestion of Professor Sat Narain Gupta, to whom I wish to express my heartfelt appreciation and eternal gratitude for his assistance and helpful criticism. He has guided me so wisely with great patience at every step of my research.

I am also indebted to Professor Greg Bell, Professor Xiaoli Gao, Professor Lakshmi S. Iyer, and Professor Haimeng Zhang for the assistance and direction they have given me in this undertaking. They are the very exemplification of kindness, integrity, and rationality. I hope our paths cross again through academic collaboration, etc.

I also extend thanks to the faculty and staff of the department of Mathematics and Statistics at UNCG led by Department Head and H. Barton Excellence Professor Shivaji, who have worked hard to provide the best possible environment for graduate students in a young doctoral program. I would like to thank the UNCG Graduate School and Professor Bell, Director of Graduate Studies of the department of Mathematics and Statistics for bestowing the Summer 2016 Research Award on me. I also appreciate everyone whom I had worked with at the UNCG's Math Help Center, led by Professor Chhetri.

Completing this work would have been all the more difficult, were it not for the support and friendship provided by Professor Ben Marlin and Professor Dani Moran at Guilford College, where I have been teaching full time since the 2016 fall semester. I am indebted to them for their help.

Lastly, I must express my gratitude to my parents and my family for their love. Without their unwavering support, I couldn't have made it through.

# TABLE OF CONTENTS

# LIST OF TABLES

Page

# LIST OF FIGURES

# CHAPTER I

# INTRODUCTION

## 1.1 Privacy Protection in Surveying Sensitive Questions

Social Desirability Bias (SDB) is the idiosyncrasy created by respondents in answering sensitive questions unfaithfully in the hope of leaving good impression on others. In addition to this Impression Management component, there also exists Self-Deception component in SDB. Some people just tend to believe that they are not engaged in socially undesirable activities and report to the interviewer accordingly, causing different kind of SDB. Paulhus (1984) recommends that Impression Management, not Self-Deception, be controlled in survey research. SDB can happen when the survey question is highly sensitive or personal. This is one of the many biases which occur during survey sampling. Other typical biases are evasive answer bias, refusal bias, nonresponse bias, selection bias, voluntary response bias, and so forth. These biases create a problem because they cause sample statistics to systematically overestimate or underestimate corresponding population parameters.

There are several techniques to promote faithful answers and to avoid Impression Management component of SDB such as the Bogus Pipeline Technique, the Unmatched Count Technique, and the Randomized Response Technique. But, in light of the principle of privacy protection, the Bogus Pipeline Technique would be a lot less desirable than other methods because it simply tricks the respondent to tell the truth and there is no privacy protected whatsoever. Privacy protection can be equated with the interviewee's power to hide his or her sensitive and personal information. Both the Unmatched Count Technique and the Randomized Response Technique adopt randomization device between the respondent and the interviewer. By placing stochastic devices between them, these two methods not only elicit truthful answers, but also protect each respondent's privacy as well.

In the following sections, we review those techniques in detail and discuss possible new applications of the Randomized Response Technique to the area of confidentiality protection in data mining and redacted datasets.

## 1.2 Various Approaches Including RRT to Circumvent SDB

### *1.2.1 Unmatched Count Technique*

This technique has a couple of different names; the Item Count Technique and the List Technique. The basic idea of the Unmatched Count Technique (UCT) is very simple. Randomly selected respondents in the control group receive a group of non-sensitive questions, and are asked to report the number of "yes" answers. After one sensitive question is added to the existing set of questions, the new set of questions is given to the other group. As members of both groups are randomly selected, we can assume that their proportions of "yes" responses towards the non-sensitive questions would be the same. Thus, we can get the unmatched count from the experimental group. As the respondents are required to simply report the number of "yes" answers, Impression Management component of SDB can be avoided. The population proportion of "yes" answers to the sensitive question can now be deduced statistically.

In many cases, it would be easy for the researcher to implement the UCT. Just paper and pencils are needed and no other complex randomization devices are required. Also, for the participants, the UCT is quite easy to understand and straightforward, providing a strong perceived sense of privacy. Studies such as Coutts and Jann (2011) and Lavender and Anderson (2009) have shown that, in practice, the UCT is more effective than other techniques because the highest perception of anonymity is found for the UCT among the respondents. However, the theory for the UTC model is not as extensive as is for the RRT models. The RRT models allow many different kinds of improvements which make these models more efficient. These improvements include optional models and two-stage models.

### *1.2.2 Bogus Pipeline Method*

The term Bogus Pipeline (BPL) was coined by Jones and Sigall (1971) to describe an imaginary dream device for psychologists, which would provide a direct pipeline to the soul. Thus, they could have access to reliable psychological indicators. Jones and Sigall (1971) proposed that respondents' answers wouldn't be contaminated by many of the biases, including SDB, if they were convinced that the device in front of them was an actual lie detector. Their explanation was that respondents didn't want to be second-guessed by a machine, trying to avoid possible loss of face while believing the true answers would be revealed regardless of their response. Roese and Jamieson (1993) showed that the BPL produced reliable effects consistent with a reduction in SDB after meta-analyzing 31 studies that had used the bogus pipeline for their research.

### *1.2.3 Randomized Response Technique*

The Randomized Response Technique (RRT) was first proposed by Warner (1965). It is a survey research method specifically designed to ask sensitive questions.

Suppose we need to estimate the proportion of drug abusers in the last 3 months in the target population. Let us have a deck of cards where 10% of the cards have the statement "I have used controlled substances without prescription at least once in the last 3 months." The rest of the cards have the statement "I have not used controlled substances without prescription in the last 3 months," written on them. The respondents are expected to give a binary answer—either "yes, this statement is correct," or "no, this statement is not correct"—to the statement on the card which they draw from the deck. Due to the randomization device—10% probability of drawing drug abuse question, the researcher has no idea of what a "yes" answer means individually or what a "no" answer means individually.

Notice that it is quite important in practice for the respondent to understand that the RRT maintains privacy, as the randomization device is invisible. Some of the respondents might not be able to grasp this probability concept easily. Without this understanding, Impression Management component of SDB cannot be overcome.

Since the RRT method was first introduced in 1965, there are many areas where the RRT models have been used. The most widely applied area for the RRT would be surveys on controlled substances and illicit drug use. Brewer (1981) estimated marijuana usage using the RRT in a population survey of Canberra, Australia and found out some paradoxical results. Akers, Massey, Clarke, and Lauer (1983) conducted a survey of teenage smoking in a Midwestern community and validated teenagers' responses by the RRT with their biochemical measure of smoking. Weissman, Steer, and Lipton (1986) conducted telephone interviews through the use of the RRT to estimate illicit drug use.

Another interesting area where the RRT plays an important role would be business management and regulatory compliance. For example, Buchman and Tracy (1982) reported a tendency toward more honest answers through the use of the RRT when they surveyed auditors' dishonest professional behavior of false sign-offs during audit procedures. Wimbush and Dalton (1997) estimated the base rate of employee theft for those personnel with access to cash, supplies, merchandise, or cash-convertible products by using two different types of the RRT. Houston and Tran (2001) conducted a mail questionnaire survey using both the RRT and the direct questioning to estimate the prevalence and type of income tax evasion. Elffers, van der Heijden, and Hezemans (2003) studied the evidence of compliance with two Dutch laws and measured self-reported compliance by use of the RRT and the adapted logistic regression. Schneider (2003) conducted an experimental study to examine whether compensation and stock ownership affect internal auditors' objectivity. In order to elicit truthful responses and overcome SDB from active internal auditors, Schneider adopted the RRT and collected randomized responses from 172 participants. It was found that stock ownership did not affect internal auditors' reporting decisions while compensation tied to stock prices made internal auditors report violations less frequently.

In the Netherlands, Lensvelt-Mulders, van der Heijden, Laudy, and van Gils (2006) validated a computer assisted RRT survey to estimate the prevalence of fraud in disability benefits. By the time of Lensvelt-Mulders et al.'s research, the actual survey to estimate the disability fraud in the Netherlands included home interviews by trained interviewers with randomized response questions. Lavender and Anderson (2009) assessed the effect of perceived anonymity on endorsements of eating disorder behaviors and attitudes among 469 undergraduate women from a university in the Northeastern United States. They used a standard anonymous true/false survey, the UCT, and the RRT. Then they compared the results generated by those three different survey techniques.

In Germany, Ostrapczuk, Musch, and Moshagen (2009) studied SDB among the highly educated and the less educated in their attitude towards foreigners, comparing their answers from direct questioning conditions and the RRT conditions. Similarly, Krumpal (2012) estimated the prevalence of xenophobia and anti–Semitism by using both the RRT as well as the direct questioning. Their results suggested that the RRT was an effective method eliciting more socially undesirable opinions and yielding more valid prevalence estimates of xenophobia and anti–Semitism than direct questioning. Also, the results indicated that with increasing topic sensitivity, the benefits of using the RRT also increased.

It is also quite common to use the RRT in order to estimate the prevalence of illegal and wrongful activities. In Hong Kong, Kwan, So, and Tam (2010) showed how truthful answers to sensitive questions about software piracy can be estimated by using the RRT. In 2011, a team of researchers at the World Anti–Doping Agency conducted their interviews with athletes at two major track and field events and surveyed how many of them had used performance enhancing drugs in the past 12 months with scrambled questions with the randomized response techniques (Rohan, 2016). Another study done by Striegel, Ulrich, and Simon (2010) estimated the prevalence of doping and illicit drug use among elite athletes by using the RRT. In the field of natural resources management, St John et al. (2011) have used the RRT to discourage environmentally harmful behaviors and estimated the proportion of farmers in north-eastern South Africa killing carnivores.

Lensvelt-Mulders, Hox, van der Heijden, and Maas (2005) discussed two meta-analyses on the RRT studies, the first on 6 individual validation studies and the second on 32 comparative studies. The authors measured the percentage of incorrect answers and found out that compared to other methods, the randomized response designs produced more valid results. They also found out that the more sensitive the topic under investigation, the higher the validity of the RRT results were.

## 1.3 Confidentiality Protection in Data Mining & Redacted Datasets

In the preceding sections, we reviewed the major characteristics and main application examples of the Randomized Response Technique in the context of privacy protection in surveying sensitive questions. But, the stochastic devices which are placed between the respondent and the interviewer in surveying sensitive questions can also be placed between collected datasets and the general public. Many governments and public organizations allow access to vast amounts of data to the general public in order to promote better decision making and to meet the needs of the various members of civil society. And they are legally bound to protect the identities of those individuals included in the specific dataset. In order to safeguard and share confidential datasets simultaneously, statisticians have developed a new field of study which is called, Statistical Disclosure Control. There are many methodologies which can strip unique identifiers to prevent data snoopers from re-identifying individuals in the released datasets. Among these are Data Aggregation, Data Swapping, Synthetic Datasets, and Adding Random Noise (Reiter & Slavkovic, 2012). Nayak, Zhang, and

Adeshiyan (2015) suggested that the Randomized Response Technique be successfully used and developed for contemporary problems like Statistical Disclosure Control.

Traditionally, the Randomized Response Technique has been employed in surveying sensitive questions to protect the respondent's privacy. As vast amount of confidential datasets have been accumulated and released, thanks to the advances in computer technologies and cheap storage costs, there exists an urgent need for confidentiality protection of redacted datasets. Nayak, Adeshiyan, and Zhang (2016) and Nayak and Adeshiyan (2016) emphasized the difference between the traditional usage of the Randomized Response Technique in privacy protection and the emerging applications of the RRT to confidentiality protection of individual identities in redacted datasets and data mining. We expect that much of the new development in the study of the Randomized Response Technique will take place in the field of confidentiality protection of data mining and redacted datasets.

### 1.4 Motivation for & Outline of the Dissertation

Chapter I has presented a brief introduction to Social Desirability Bias and discussed several techniques to promote faithful answers to sensitive questions. It has also discussed how those techniques were applied in practice. Furthermore, we compared the traditional role which the RRT has played in privacy protection with the emerging application in confidentiality protection and privacy-preserving in data mining.

Chapter II presents four foundational studies in the RRT field and corresponding models including both binary and quantitative models, which serve as the basis for the proposed models in this dissertation.

Chapter III discusses the traditional split-sample based binary RRT models. The first one is based on Greenberg's unrelated question model and the other on Warner's indirect question model.

Chapter IV presents the latest advances in the binary RRT area with two-question approach. Instead of using the traditional split-sample approach, we introduce the two-question approach to estimate the level of sensitivity. By choosing appropriate value of the two-stage parameter $T$, the proposed model achieves smaller estimator variance than the competing model.

Chapter V presents how the simulations are set up and discusses the results of simulations for the comparison of the estimators. We also present the results of the suitable $T$ intervals by using `hit or miss` simulation methods.

Chapter VI presents the concluding remarks of this dissertation and possible future work.

Appendix A & B present the R program codes for the simulations on comparing performances of the various estimators.

# CHAPTER II
# VARIOUS RRT MODELS

## 2.1 Indirect Question RRT

The basic idea of the Randomized Response Technique is rather simple. If a randomization mechanism can be placed between the interviewee and the interviewer, who thus is not allowed to distinguish the meaning of each individual response, then increased level of privacy will be ensured. This increased level of privacy will facilitate increased level of cooperation and elicit a more truthful answer from the interviewee. Warner (1965) pioneered this very interesting idea of putting a randomization device to deal with evasive answer bias, especially associated with those personal or controversial survey questions.

As in Warner (1965), it is assumed likewise that a simple random sample is drawn with replacement from the population throughout this dissertation. In this section, we show two models of the indirect question RRT; one for the binary response (Warner, 1965) and another for the quantitative response (Warner, 1971).

### *2.1.1 Binary Indirect Question RRT*

Warner (1965) proposed a spinner with probability $p$ pointing to the letter A and with probability $(1-p)$ pointing to the letter B. Every respondent belongs to either Group A (the sensitive group) or Group B (the non-sensitive group). The spinner is run without the interviewer's presence and the interviewee is to report a "Yes" or a "No" to indicate whether or not the group the spinner is pointing to is the group he or she actually belongs to.

Let $P_y$ be the probability of a "Yes" response from a respondent. Note that a "Yes" response can be provided in two ways. One is when the respondent belongs to Group A while the spinner points to A. Another is when he or she belongs to Group B while the spinner points to B. Let $\pi$ be the proportion of a population that belongs to Group A. We want to estimate $\pi$.

Then $P_y$ can be expressed as follows.

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

Solving for $\pi$, we have

$$\pi = \frac{P_y - (1 - p)}{2p - 1}.$$

Thus, the Warner's estimate of $\pi$ is given by

$$\widehat{\pi}_w = \frac{\widehat{P_y} - (1 - p)}{2p - 1} \qquad \left(p \neq \frac{1}{2}\right), \tag{2.2}$$

where $\widehat{P_y}$ is the proportion of "Yes" responses in the survey.

Notice that $\widehat{P_y}$ is an unbiased estimator as well as the Maximum Likelihood Estimator (MLE) of $P_y$. Taking expected value on Equation (2.2), we get

$$E\left(\widehat{\pi}_w\right) = \frac{E\left(\widehat{P_y}\right) - (1 - p)}{2p - 1} = \frac{P_y - (1 - p)}{2p - 1} = \pi.$$

Thus, $\widehat{\pi}$ from Equation (2.2) is an unbiased estimator of $\pi$.

The variance of $\widehat{\pi}$ is given by

$$Var\left(\widehat{\pi}_w\right) = \frac{1}{(2p-1)^2} Var\left(\widehat{P_y}\right) \tag{2.3}$$

$$= \frac{1}{(2p-1)^2} \left\{ \frac{P_y(1-P_y)}{n} \right\}. \tag{2.4}$$

after using $Var(\widehat{P_y}) = P_y(1-P_y)/n$. On substituting $P_y$ from Equation (2.1) into Equation (2.4), we have the variance of the Warner's estimator as given by

$$Var(\widehat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{2.5}$$

with

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

The second term in Equation (2.6) is the penalty for using the RRT and will go down as $n$ increases. It is advised that one should select $p$ away from $\frac{1}{2}$ as shown in Equation (2.2), and close to 0 or 1.

### *2.1.2 Quantitative Indirect Question RRT*

The quantitative models of the indirect question Randomized Response Technique was first proposed by Warner (1971) in a rather passing manner without any formal presentation.

First, Warner (1971) briefly discussed the additive model, saying "One Obvious approach is to specify that the reported $Z$ would be the true $X$ plus a random term $Y$ from a known distribution, the random term $Y$ being selected and added in private." Gupta, Shabbir, and Sehra (2010) later generalized this first additive model and formally presented the two-stage optional additive model of the indirect question RRT. One can verify this by substituting $\omega = 1$ and $T = 1$ for the generalized model of Gupta, Shabbir, and Sehra (2010).

Second, Warner (1971) also proposed the multiplicative model right after the preceding sentence, adding "A related model could specify that the reported $Z$ would be the true $X$ multiplied by a random term $Y$ from a known distribution, the multiplier $Y$ being selected and the multiplication accomplished in private." This second suggested model would be later presented in a formal manner by Eichhorn and Hayre (1983). They also showed that this model is generally superior to the quantitative unrelated question RRT model which is described in Subsection 2.2.2.

Let us briefly talk about these two quantitative models which were originally suggested by Warner (1971).

#### *2.1.2.1 Additive Model*

The first quantitative model of Warner (1971) is the additive model which is quite trivial as shown below. Let $\mu_y$ and $\sigma_y^2$ respectively be the known mean and variance from a known distribution, and $\mu_x$ and $\sigma_x^2$ respectively be the unknown mean and variance of the sensitive question in the population. Assume random variables $X$ and $Y$ are independent. Let $Z$ be the reported response from a respondent. Then $Z$ can be expressed as

$$Z = X + Y$$

with

$$E(Z) = \mu_z = \mu_x + \mu_y, \tag{2.7}$$

$$Var(Z) = Var(X) + Var(Y)$$

$$= \sigma_x^2 + \sigma_y^2. \tag{2.8}$$

Solving Equation (2.7) for $\mu_x$, we have

$$\mu_x = \mu_z - \mu_y.$$

This leads to the following estimator,

$$\widehat{\mu}_x = \overline{Z} - \mu_y, \tag{2.9}$$

where $\overline{Z}$ is the sample mean of the quantitative responses in the survey.

It is easy to verify that $\widehat{\mu}_x$ is an unbiased estimator with its variance given by

$$Var\left(\widehat{\mu}_x\right) = \frac{1}{n} Var\left(Z\right) = \frac{1}{n}\left(\sigma_x^2 + \sigma_y^2\right). \tag{2.10}$$

#### *2.1.2.2 Multiplicative Model*

The second model of Warner (1971) was the multiplicative model, which later was formally presented by Eichhorn and Hayre (1983). Let $X$ denote the answer to the sensitive question with unknown $\mu_x$ and $\sigma_x^2$. Let $Y$ be a random variable independent of $X$ with known $\mu_y$ and $\sigma_y^2$. Assume also $X \geq 0$ and $Y > 0$. Let $Z$ be the reported response from a respondent. Then $Z$ can be expressed as

$$Z = X \times Y$$

with

$$E(Z) = \mu_z = E(XY) = E(X)E(Y) = \mu_x \mu_y, \tag{2.11}$$

$$\begin{aligned} Var(Z) &= E(X^2)E(Y^2) - (E(X))^2 (E(Y))^2 \\ &= \left(\sigma_x^2 + \mu_x^2\right)\left(\sigma_y^2 + \mu_y^2\right) - \mu_x^2 \mu_y^2 \\ &= \left(\sigma_x^2 + \mu_x^2\right)\sigma_y^2 + \left(\sigma_x^2 + \mu_x^2\right)\mu_y^2 - \mu_x^2 \mu_y^2 \\ &= \left(\sigma_x^2 + \mu_x^2\right)\sigma_y^2 + \sigma_x^2 \mu_y^2. \end{aligned} \tag{2.12}$$

In this setup, the multiplicative model can compromise anonymity because a non-zero reported response $Z$ means $X$ must be non-zero, which clearly indicates some degree of sensitive behavior is taking place.

Solving Equation (2.11) for $\mu_x$, we have

$$\mu_x = \frac{\mu_z}{\mu_y}.$$

This leads to the following estimator,

$$\widehat{\mu}_x = \frac{\overline{Z}}{\mu_y}, \tag{2.13}$$

where $\overline{Z}$ is the sample mean of the quantitative responses in the survey. It is easy to verify that $\widehat{\mu}_x$ is an unbiased estimator with its variance given by

$$Var\left(\widehat{\mu}_x\right) = \frac{1}{n\mu_y^2} Var\left(Z\right) = \frac{1}{n\mu_y^2}\left\{\left(\sigma_x^2 + \mu_x^2\right)\sigma_y^2 + \sigma_x^2\mu_y^2\right\} = \frac{1}{n}\left\{\left(\sigma_x^2 + \mu_x^2\right)\left(\frac{\sigma_y}{\mu_y}\right)^2 + \sigma_x^2\right\}. \tag{2.14}$$

## 2.2 Unrelated Question RRT

The unrelated question model was first proposed by Greenberg, Abul-Ela, Simmons, and Horvitz (1969). It was the binary model and soon after, Greenberg, Kuebler, Abernathy, and Horvitz (1971) introduced the quantitative version of the Unrelated Question RRT. The main idea behind these models was that, by adding innocuous questions to the questionnaire with a pre-assigned probability, the response would be scrambled and the researcher could protect the respondent's privacy to elicit truthful answers. In this section, we first discuss two binary unrelated question models; one with known prevalence of the innocuous characteristic and another with unknown prevalence of the innocuous characteristic. Then we examine the quantitative unrelated question model later in the section.

The main advantage of the unrelated question RRT model is that some of the respondents answer a non-sensitive question, thereby increasing cooperation from the respondents.

### *2.2.1 Binary Unrelated Question RRT*

In this binary model, a randomization device is used to ask a respondent the sensitive binary question with pre-assigned probability $p_a$ as well as an innocuous question (whose prevalence is already known) with probability $(1 - p_a)$.

Let $\pi_a$ be the known prevalence of the unrelated characteristic and $\pi$ be the unknown prevalence of the sensitive characteristic. Let $P_y$ be the probability of a "Yes" response from a respondent. Then $P_y$ can be expressed as

$$P_y = p_a \pi + (1 - p_a)\pi_a. \tag{2.15}$$

Solving for $\pi$, we have

$$\pi = \frac{P_y - (1 - p_a)\pi_a}{p_a}.$$

This leads to the estimator of Greenberg et al. (1969)

$$\widehat{\pi} = \frac{\widehat{P_y} - (1 - p_a)\pi_a}{p_a}, \tag{2.16}$$

where $\widehat{P_y}$ is the proportion of "Yes" responses in the survey. It is easy to show that $\widehat{\pi}$ is an unbiased estimator with its variance given by

$$Var\,(\widehat{\pi}) = \frac{P_y(1-P_y)}{np_a^2}. \tag{2.17}$$

### 2.2.2 Binary Unrelated Question RRT with Unknown $\pi_a$

The model described in this subsection is almost identical to the model in the preceding subsection. The only difference is that we no longer assume we somehow know the prevalence of the innocuous characteristic. Instead, we treat it as unknown. A randomization device is used to ask a respondent the sensitive binary question with pre-assigned probability $p_i$ as well as an innocuous question (whose prevalence is unknown) with probability $(1-p_i)$.

Let $\pi_a$ be the unknown prevalence of the unrelated characteristic and $\pi$ be the unknown prevalence of the sensitive characteristic. Let $P_y$ be the probability of a "Yes" response from a respondent. Then $P_y$ can be expressed as

$$P_y = p\pi + (1-p)\pi_a. \tag{2.18}$$

Equation (2.18) can be rearranged as

$$P_y - p\pi = (1-p)\pi_a. \tag{2.19}$$

As Equation (2.19) includes two parameters ($\pi$ and $\pi_a$), it cannot be handled with one set of responses. Assume we have two independent samples with sizes $n_1$ and $n_2$ respectively ($n_1 + n_2 = n$). Let us also assume that $p_1$ and $p_2$ are two different pre-assigned probabilities associated with the different randomization devices used in the two samples. Using Equation (2.19) for the two independent samples, we have

$$P_{y_1} - p_1\pi = (1-p_1)\pi_a \qquad \text{and} \qquad P_{y_2} - p_2\pi = (1-p_2)\pi_a. \tag{2.20}$$

With $\lambda = \frac{(p_1-1)}{(p_2-1)}$ as in Greenberg et al. (1969), we have

$$\pi = \frac{\lambda P_{y_2} - P_{y_1}}{\lambda p_2 - p_1}. \tag{2.21}$$

From Equation (2.21), we have the estimator for $\pi$ as

$$\widehat{\pi} = \frac{\lambda \widehat{P_{y_2}} - \widehat{P_{y_1}}}{\lambda p_2 - p_1}, \tag{2.22}$$

where $\widehat{P_{y_1}}$ and $\widehat{P_{y_2}}$ are the proportions of "Yes" responses in the two samples with sample size $n_1$ and $n_2$ respectively. Note that $\widehat{\pi}$ is unbiased as shown below.

$$E\left(\widehat{\pi}\right) = \frac{\lambda E\left(\widehat{P_{y_2}}\right) - E\left(\widehat{P_{y_1}}\right)}{\lambda p_2 - p_1} = \frac{\lambda P_{y_2} - P_{y_1}}{\lambda p_2 - p_1} = \pi. \tag{2.23}$$

Using $Var(\widehat{P_{y_1}}) = P_{y_1}(1-P_{y_1})/n_1$ and $Var(\widehat{P_{y_2}}) = P_{y_2}(1-P_{y_2})/n_2$, the variance of $\widehat{\pi}$ is

$$Var(\widehat{\pi}) = \frac{1}{(\lambda p_2 - p_1)^2}\{\lambda^2 Var(\widehat{P_{y_2}}) + Var(\widehat{P_{y_1}})\}$$

$$= \frac{1}{(\lambda p_2 - p_1)^2}\left\{\lambda^2 \frac{P_{y_2}(1-P_{y_2})}{n_2} + \frac{P_{y_1}(1-P_{y_1})}{n_1}\right\}. \tag{2.24}$$

Notice that the two samples are independent so that the covariance term does not exist in Equation (2.24). Using $n_1 = n - n_2$, we can rewrite Equation (2.24) as

$$Var(\widehat{\pi}) = \frac{1}{(\lambda p_2 - p_1)^2} \left\{ \lambda^2 \frac{P_{y_2}(1 - P_{y_2})}{n_2} + \frac{P_{y_1}(1 - P_{y_1})}{n - n_2} \right\}. \tag{2.25}$$

After taking partial derivative of Equation (2.25), the optimal ratio of $\frac{n_1}{n_2}$, which gives the minimum variance, can be obtained:

$$\frac{\partial Var(\widehat{\pi})}{\partial n_2} = \frac{1}{(\lambda p_2 - p_1)^2} \frac{\partial}{\partial n_2} \left( -\lambda^2 \frac{P_{y_2}(1 - P_{y_2})}{n_2^2} + \frac{P_{y_1}(1 - P_{y_1})}{(n - n_2)^2} \right) = 0. \tag{2.26}$$

Solving Equation (2.26) for $\frac{n_1}{n_2}$, the optimal ratio of $\left(\frac{n_1}{n_2}\right)_{opt(\widehat{\pi})}$ will be

$$\left(\frac{n_1}{n_2}\right)_{opt(\widehat{\pi})} = \frac{1}{\lambda} \sqrt{\frac{P_{y_1}(1 - P_{y_1})}{P_{y_2}(1 - P_{y_2})}} = \frac{(1 - p_2)}{(1 - p_1)} \sqrt{\frac{P_{y_1}(1 - P_{y_1})}{P_{y_2}(1 - P_{y_2})}}. \tag{2.27}$$

Let us solve Equations (2.20) for $\pi_a$. We have

$$p_2 P_{y_1} - p_1 P_{y_2} = (p_2 - p_1)\pi_a. \tag{2.28}$$

Solving Equation (2.28) for $\pi_a$, we have

$$\pi_a = \frac{p_2 P_{y_1} - p_1 P_{y_2}}{p_2 - p_1}. \tag{2.29}$$

By replacing $P_{y_1}$ and $P_{y_2}$ with their unbiased MLEs, the estimator for $\pi_a$ is

$$\widehat{\pi}_a = \frac{p_2\widehat{P_{y_1}} - p_1\widehat{P_{y_2}}}{p_2 - p_1}. \tag{2.30}$$

It is easy to show that $\widehat{\pi}_a$ is unbiased for $\pi_a$ with its variance given by

$$Var\left(\widehat{\pi}_a\right) = \frac{1}{(p_2 - p_1)^2}\left\{{p_2}^2\left\{\frac{P_{y_1}(1 - P_{y_1})}{n_1}\right\} + {p_1}^2\left\{\frac{P_{y_2}(1 - P_{y_2})}{n_2}\right\}\right\}. \tag{2.31}$$

### *2.2.3 Quantitative Unrelated Question RRT*

Very much like the binary response models in previous subsections, the researcher in this quantitative model will also ask a sensitive question with pre-assigned probability $p_a$ and an innocuous question with probability $(1 - p_a)$.

Let $\mu_y$ and $\sigma_y^2$ be the known mean and variance of an unrelated question. And let $\mu_x$ and $\sigma_x^2$ be the unknown mean and variance of the sensitive question in the population. Let $Z$ be the reported response from a respondent. Then $Z$ can be expressed as

$$Z = \begin{cases} X & \text{with probability } p_a \text{ (sensitive question),} \\ Y & \text{with probability } (1 - p_a) \text{ (non-sensitive question),} \end{cases}$$

with

$$E(Z) = \mu_z = p_a \mu_x + (1 - p_a)\mu_y, \tag{2.32}$$

$$Var(Z) = p_a E\left(X^2\right) + (1 - p_a) E\left(Y^2\right) - \mu_z^2$$

$$= p_a\left(\sigma_x^2 + \mu_x^2\right) + (1 - p_a)\left(\sigma_y^2 + \mu_y^2\right) - \mu_z^2. \tag{2.33}$$

Solving Equation (2.32) for $\mu_x$, we have

$$\mu_x = \frac{\mu_z - (1 - p_a)\mu_y}{p_a}.$$

This leads to the estimator of Greenberg et al. (1971),

$$\widehat{\mu}_x = \frac{\overline{Z} - (1 - p_a)\mu_y}{p_a}, \tag{2.34}$$

where $\overline{Z}$ is the sample mean of the quantitative responses in the survey. It is easy to show that $\widehat{\mu}_x$ is an unbiased estimator with its variance given by

$$Var\left(\widehat{\mu}_x\right) = \frac{1}{np_a^2} Var\left(Z\right) = \frac{1}{np_a^2}\left(\sigma_y^2 + p_a(\sigma_x^2 - \sigma_y^2) + p_a(1 - p_a)(\mu_x - \mu_y)^2\right). \tag{2.35}$$

## 2.3 Two-Stage RRT

Mangat and Singh (1990) introduced a Two-Stage RRT model by injecting an element of truthful responses into the indirect binary question randomized response model of Warner (1965).

In order to have more truthful answers, they placed one more randomization device into the original Warner's model. The first randomization device has two options: (1) 'Do you belong to Group A?', and (2) 'Go to the second randomization device,' And, the second stage—or the second randomization device—is nothing but the Warner's randomization device. The pre-assigned probabilities of two options (1) and (2) are $T$ and $(1 - T)$, respectively. Because the entire process remains unobserved by the interviewer, as in the Warner's model, the interviewee can maintain privacy regardless of the answer either from the first randomization device or from the Warner's randomization device.

Let $P_y$ be the probability of a "Yes" response from a respondent under this model. $P_y$ is given by

$$P_y = T\pi + (1-T)\{\pi p + (1-\pi)(1-p)\} = \{T + (2p-1)(1-T)\}\pi + (1-T)(1-p). \qquad (2.36)$$

Rewriting this equation for $\pi$, we have

$$\pi = \frac{P_y - (1-p)(1-T)}{T + (2p-1)(1-T)} = \frac{P_y - (1-p)(1-T)}{(2p-1) + 2T(1-p)}.$$

This leads to the Mangat and Singh's estimator for $\pi$, given by

$$\widehat{\pi}_m = \frac{\widehat{P_y} - (1-p)(1-T)}{(2p-1) + 2T(1-p)}. \tag{2.37}$$

where $\widehat{P_y}$ is the proportion of "Yes" responses in the survey. As $\widehat{P_y}$ is both unbiased and the MLE of $P_y$, $\widehat{\pi}_m$ is unbiased too. This can be verified from the following.

$$E\left(\widehat{\pi}_m\right) = \frac{E\left(\widehat{P_y}\right) - (1-p)(1-T)}{(2p-1) + 2T(1-p)} = \frac{P_y - (1-p)(1-T)}{(2p-1) + 2T(1-p)} = \pi.$$

Also the variance of the estimator is given by

$$Var(\widehat{\pi}_m) = \frac{1}{\{(2p-1) + 2T(1-p)\}^2} Var(\widehat{P_y}) \tag{2.38}$$

$$= \frac{1}{\{(2p-1) + 2T(1-p)\}^2} \left\{ \frac{P_y(1-P_y)}{n} \right\}. \tag{2.39}$$

Using Equation (2.36), this can be rewritten as

$$Var(\widehat{\pi}_m) = \frac{\pi(1-\pi)}{n} + \frac{(1-T)(1-p)\{1-(1-T)(1-p)\}}{n\{(2p-1) + 2T(1-p)\}^2} \tag{2.40}$$

with

$$\widehat{Var}(\widehat{\pi}_m) = \frac{\widehat{\pi}_m(1-\widehat{\pi}_m)}{n-1} + \frac{(1-T)(1-p)\{1-(1-T)(1-p)\}}{n\{(2p-1) + 2T(1-p)\}^2}. \tag{2.41}$$

Mangat and Singh (1990) also showed that $Var(\widehat{\pi}_m)$, the variance of their estimator, is smaller than $Var(\widehat{\pi}_w)$ of Warner (1965), when $T > \frac{1-2p}{1-p}$. As $\frac{1-2p}{1-p} < 1$ for $0 < p < 1$, we can always assign a meaningful value of $T$ between $\frac{1-2p}{1-p}$ and 1.

## 2.4 Optional RRT

It is reasonable to assume that some proportion of the population might not feel the survey question is sensitive and would give candid answers if they get the option to answer truthfully. Instead of injecting an element of truth by the researchers, as in the Two-Stage Model of Mangat and Singh (1990), we can incorporate this unknown proportion of truthfulness in a different manner into a new model. In this Optional Model, the respondent has the freedom to choose how to answer the question. If the respondent feels the question is sensitive, he or she can give a scrambled response. If the respondent doesn't feel it's a sensitive question, he or she can just give a true answer. This optional randomization process takes place without being observed by the researcher, who has no idea of what method the respondent chose and what a "Yes" response means.

In the Two-Stage Model of Section 2.3, the truth parameter $T$ is pre–assigned by the interviewer, thus was a known constant prior to using the two randomization devices. In this Optional Model, the sensitivity level ($\omega$) of a specific question is defined to be the population proportion of subjects who feel the question is sensitive. Notice that there are two unknown parameters in this model ($\pi$ and $\omega$). The Optional Randomized Response models were first proposed by Gupta (2001) for the binary case, and by Gupta, Gupta, and Singh (2002) for the quantitative case. The characteristics of the models have been discussed in great depth by Gupta and Thornton (2002), Gupta and Shabbir (2004), Gupta, Thornton, Shabbir, and Singhal (2006), Gupta, Shabbir, and Sehra (2010), and Gupta, Mehta, Shabbir, and Dass (2013).

In this section, the binary optional model of Gupta (2001) is discussed first and the quantitative optional model of Gupta, Gupta, and Singh (2002) is examined next.

### *2.4.1 Binary Optional RRT*

The probability of a "Yes" response in this model can be expressed as

$$P_y = (1-\omega)\pi + \omega\{\pi p + (1-\pi)(1-p)\}. \tag{2.42}$$

Equation (2.42) can be rearranged as

$$P_y - \pi = (p-1)(2\pi-1)\omega. \tag{2.43}$$

As Equation (2.43) includes two parameters ($\pi$ and $\omega$), it cannot be handled with one set of responses. Assume we have two independent samples with sizes $n_1$ and $n_2$ respectively ($n_1 + n_2 = n$). Let us also assume that $p_1$ and $p_2$ are different probabilities associated with the different Warner's devices used in the two samples. Using Equation (2.43) for the two independent samples, we have

$$P_{y_1} - \pi = (p_1-1)(2\pi-1)\omega \qquad \text{and} \qquad P_{y_2} - \pi = (p_2-1)(2\pi-1)\omega. \tag{2.44}$$

With $\lambda = (p_1-1)/(p_2-1)$ as in Greenberg et al. (1969), we have

$$\pi = \frac{\lambda P_{y_2} - P_{y_1}}{\lambda - 1}. \tag{2.45}$$

From Equation (2.45), we have the Gupta estimator for $\pi$ as

$$\widehat{\pi}_g = \frac{\lambda \widehat{P_{y_2}} - \widehat{P_{y_1}}}{\lambda - 1}. \tag{2.46}$$

where $\widehat{P_{y_1}}$ and $\widehat{P_{y_2}}$ are the proportions of "Yes" responses in the two samples with sample size $n_1$ and $n_2$ respectively. Note that $\widehat{\pi}_g$ is unbiased as shown below:

$$E\left(\widehat{\pi}_g\right) = \frac{\lambda E\left(\widehat{P_{y_2}}\right) - E\left(\widehat{P_{y_1}}\right)}{\lambda - 1} = \frac{\lambda P_{y_2} - P_{y_1}}{\lambda - 1} = \pi. \tag{2.47}$$

Using $Var(\widehat{P_{y_1}}) = P_{y_1}(1 - P_{y_1})/n_1$ and $Var(\widehat{P_{y_2}}) = P_{y_2}(1 - P_{y_2})/n_2$, the variance of $\widehat{\pi}_g$ is

$$Var(\widehat{\pi}_g) = \frac{1}{(\lambda - 1)^2}\{\lambda^2 Var(\widehat{P_{y_2}}) + Var(\widehat{P_{y_1}})\}$$

$$= \frac{1}{(\lambda - 1)^2}\left\{\lambda^2 \frac{P_{y_2}(1 - P_{y_2})}{n_2} + \frac{P_{y_1}(1 - P_{y_1})}{n_1}\right\}. \tag{2.48}$$

Notice that the two samples are independent so that the covariance term does not exist in Equation (2.48). Using $n_1 = n - n_2$, we can rewrite Equation (2.48) as

$$Var(\widehat{\pi}_g) = \frac{1}{(\lambda - 1)^2}\left\{\lambda^2 \frac{P_{y_2}(1 - P_{y_2})}{n_2} + \frac{P_{y_1}(1 - P_{y_1})}{n - n_2}\right\} \tag{2.49}$$

After taking partial derivative of Equation (2.49), the optimal ratio of $\frac{n_1}{n_2}$, which gives the minimum variance, can be obtained:

$$\frac{\partial Var(\widehat{\pi}_g)}{\partial n_2} = \frac{1}{(\lambda-1)^2}\frac{\partial}{\partial n_2}\left(-\lambda^2\frac{P_{y_2}(1-P_{y_2})}{n_2^2} + \frac{P_{y_1}(1-P_{y_1})}{(n-n_2)^2}\right) = 0. \tag{2.50}$$

Solving Equation (2.50) for $\frac{n_1}{n_2}$, the optimal ratio of $\left(\frac{n_1}{n_2}\right)_{opt(\widehat{\pi}_g)}$ will be

$$\left(\frac{n_1}{n_2}\right)_{opt(\widehat{\pi}_g)} = \frac{1}{\lambda}\sqrt{\frac{P_{y_1}(1-P_{y_1})}{P_{y_2}(1-P_{y_2})}} = \frac{(1-p_2)}{(1-p_1)}\sqrt{\frac{P_{y_1}(1-P_{y_1})}{P_{y_2}(1-P_{y_2})}}. \tag{2.51}$$

Let us solve Equations (2.44) for $\omega$. We have

$$P_{y_1} - P_{y_2} = (p_1 - p_2)(2\pi - 1)\omega. \tag{2.52}$$

Solving Equation (2.52) for $\omega$ and substituting $\pi = \frac{\lambda P_{y_2} - P_{y_1}}{\lambda - 1}$ and $\lambda = \frac{(p_1-1)}{(p_2-1)}$ from Equations (2.45), we have

$$\omega = \frac{P_{y_1} - P_{y_2}}{2P_{y_1}(1-p_2) - 2P_{y_2}(1-p_1) - (p_1-p_2)}. \tag{2.53}$$

By replacing $P_{y_1}$ and $P_{y_2}$ with their unbiased MLEs, the Gupta estimator for $\omega$ is

$$\widehat{\omega}_g = \frac{\widehat{P_{y_1}} - \widehat{P_{y_2}}}{2\widehat{P_{y_1}}(1-p_2) - 2\widehat{P_{y_2}}(1-p_1) - (p_1-p_2)}. \tag{2.54}$$

Given the fact that $\widehat{\omega}_g$ is a ratio of combinations of two random variables, calculation of its mean and variance will require some approximation. Sihm and

Gupta (2015) showed, up to first–order Taylor approximation, $\widehat{\omega}_g$ is an unbiased estimator for $\omega$ with its variance given by

$$Var\left(\widehat{\omega}_g\right) \approx \frac{(p_1-p_2)^2\left\{(2P_{y_2}-1)^2\left\{\frac{P_{y_1}(1-P_{y_1})}{n_1}\right\}+(2P_{y_1}-1)^2\left\{\frac{P_{y_2}(1-P_{y_2})}{n_2}\right\}\right\}}{\{(1-p_2)(2P_{y_1}-1)-(1-p_1)(2P_{y_2}-1)\}^4}. \tag{2.55}$$

### *2.4.2 Quantitative Optional RRT*

The quantitative optional RRT model was first proposed by Gupta, Gupta, and Singh (2002). It was an improvement on the multiplicative model of Eichhorn and Hayre (1983) with the option of answering the sensitive question directly without any randomization device if the respondent would feel comfortable about doing so. Gupta, Gupta, and Singh (2002) showed that the new estimator is more efficient than the Eichhorn and Hayre (1983) estimator.

Each respondent selected chooses one of the following two options. (1) The respondent reports the truthful response $X$ directly if the respondent feels comfortable about doing so, and (2) The respondent reports the scrambled response $SX$, where $S$ denotes the independent scrambling variable with $\mu_s = 1$ and known $\sigma_s$. Let $Z$ be the reported response from a respondent. Assuming both random variables $X$ and $S$ are positive valued, the model is given by

$$Z = S^Y X, \tag{2.56}$$

where $Y$ is a random variable defined as

$$Y = \begin{cases} 1 & \text{with unknown probability } \omega, \\ 0 & \text{with unknown probability } (1-\omega). \end{cases} \tag{2.57}$$

Let $\mu_x$ and $\sigma_x^2$ respectively be the unknown mean and variance of the sensitive question in the population. We can show that

$$E(Z) = E(S^Y X) = E(S^Y X \mid Y=1)P(Y=1) + E(S^Y X \mid Y=0)P(Y=0)$$

$$= E(SX)P(Y=1) + E(X)P(Y=0)$$

$$= E(S)E(X)\omega + E(X)(1-\omega)$$

$$= \mu_x\omega + \mu_x(1-\omega) = \mu_x. \tag{2.58}$$

Thus, an unbiased estimator is given by

$$\widehat{\mu}_x = \bar{Z}, \tag{2.59}$$

with

$$Var\left(\widehat{\mu}_x\right) = \frac{1}{n}\left\{\left(\sigma_x^2 + \mu_x^2\right)\sigma_y^2\,\omega + \sigma_x^2\right\}. \tag{2.60}$$

It is easy to show that the variance in Equation (2.60) is smaller than the variance in Equation (2.14) because $0 \leq \omega \leq 1$. Gupta, Gupta, and Singh (2002) proposed an estimator for $\omega$, up to first-order Taylor approximation, given by

$$\widehat{\omega} \approx \frac{\frac{1}{n}\sum \log(Z_i) - \log\left(\frac{1}{n}\sum Z_i\right)}{E(\log(S))}, \tag{2.61}$$

with

$$\widehat{Var}\,(\widehat{\omega}) \approx \frac{\widehat{\omega}(1-\widehat{\omega})}{n-1}. \tag{2.62}$$

## 2.5 Conclusion

In this chapter, we presented four foundational studies in the RRT field and corresponding models including both binary and quantitative models. For several decades, these models have been developed to improve the efficiency of the RRT models. The binary models in this chapter will be used as the building blocks for the later models in Chapter III and IV.

# CHAPTER III

# MODIFIED BINARY OPTIONAL RRT MODELS WITH SPLIT-SAMPLE APPROACH

## 3.1 Introduction

Greenberg et al. (1969) used multiple sub-samples for estimating multiple parameters associated with the RRT methods. We discuss here two methods that have been particularly developed for binary RRT models to estimate the prevalence of a sensitive characteristic and the sensitivity level of the underlying research question. The first one is based on the unrelated binary question RRT model of Greenberg et al. (1969). It was developed by Gupta, Tuck, Spears Gill, and Crowe (2013) . The second split-sample method is based on the Warner's indirect question RRT model and was proposed by Sihm and Gupta (2015).

In this chapter, major characteristics of these two models are discussed. And their estimators are proposed and variances are derived up to first-order Taylor approximation. A feature that renders the split-sample approach less acceptable is its relatively low efficiency. In most of the cases, the split-sample approach requires bigger sample size to achieve a level of efficiency associated with other models. This leads us to another method of estimating multiple parameters in the next chapter.

## 3.2 Binary Optional Unrelated RRT

This model was proposed by Gupta, Tuck, Spears Gill, and Crowe (2013) as a generalization of the original Greenberg et al. (1969, 1971) unrelated question models by giving respondents the option of responding to the sensitive question directly if they consider the question non-sensitive, while they can still give scrambled response by using the Greenberg et al. (1969) model for binary response and by using the Greenberg et al. (1971) model for quantitative response, if they feel the question is sensitive.

Let $\pi_a$ be the known prevalence of an unrelated characteristic, $\pi$ be the unknown prevalence of the sensitive characteristic, $p$ be the pre-assigned probability of the respondent selecting the sensitive question, and $\omega$ be the unknown sensitivity level of the survey question in the population. Sensitivity level means the proportion of respondents in the population who would consider the question sensitive and subsequently opt to use a randomization device.

The probability of a "Yes" response ($P_y$) in this model can be expressed as

$$P_y = (1-\omega)\pi + \omega\{\pi p + (1-p)\pi_a\}. \tag{3.1}$$

Equation (3.1) can be rearranged as

$$P_y - \pi = \omega(1-p)(\pi_a - \pi). \tag{3.2}$$

Using two independent samples with sizes $n_1$ and $n_2$ respectively, and assuming that $p_1$ and $p_2$ are two different pre-assigned probabilities of the respondents selecting the sensitive question in the two samples, Equation (3.2) can be written as

$$P_{y_1} - \pi = \omega(1-p_1)(\pi_a - \pi) \qquad \text{and} \qquad P_{y_2} - \pi = \omega(1-p_2)(\pi_a - \pi). \tag{3.3}$$

Solving for $\pi$, we have

$$\pi = \frac{\lambda P_{y_2} - P_{y_1}}{\lambda - 1} \qquad (p_1 \neq 1, p_2 \neq 1, p_1 \neq p_2, \pi_a \neq \pi), \qquad \text{where } \lambda = \frac{(p_1 - 1)}{(p_2 - 1)}. \tag{3.4}$$

Equation (3.4) leads to the unbiased estimator of Gupta, Tuck, Spears Gill, and Crowe (2013) for $\pi$, given by

$$\widehat{\pi}_{gu} = \frac{\lambda \widehat{P_{y_2}} - \widehat{P_{y_1}}}{\lambda - 1}, \tag{3.5}$$

with its variance given by

$$Var(\widehat{\pi}_{gu}) = \frac{1}{(\lambda - 1)^2} \left\{ \lambda^2 \frac{P_{y_2}(1 - P_{y_2})}{n_2} + \frac{P_{y_1}(1 - P_{y_1})}{n_1} \right\}. \tag{3.6}$$

Similarly from Equations (3.3), we have

$$P_{y_1} - P_{y_2} = \omega(p_2 - p_1)(\pi_a - \pi). \tag{3.7}$$

Solving Equation (3.7) for $\omega$ and substituting $\pi = \frac{\lambda P_{y_2} - P_{y_1}}{\lambda - 1}$ and $\lambda = \frac{(p_1 - 1)}{(p_2 - 1)}$ from Equations (3.4), we have,

$$\omega = \frac{P_{y_1} - P_{y_2}}{(p_2 - p_1)\pi_a + (1 - p_2)P_{y_1} - (1 - p_1)P_{y_2}} \qquad (p_1 \neq p_2, \pi_a \neq \pi), \tag{3.8}$$

which leads to an estimator of $\omega$ given by

$$\widehat{\omega}_{gu} = \frac{\widehat{P_{y_1}} - \widehat{P_{y_2}}}{(p_2 - p_1)\pi_a + (1-p_2)\widehat{P_{y_1}} - (1-p_1)\widehat{P_{y_2}}}. \tag{3.9}$$

Gupta, Tuck, Spears Gill, and Crowe (2013) approximate the mean and variance of this estimator by using first-order Taylor expansion. We first rewrite $\widehat{\omega}_{gu}$ as

$$\widehat{\omega}_{gu} \approx \widehat{\omega}(P_{y_1}, P_{y_2}) + \left.\frac{\partial \widehat{\omega}(\widehat{P_{y_1}}, \widehat{P_{y_2}})}{\partial \widehat{P_{y_1}}}\right|_{P_{y_1}, P_{y_2}} (\widehat{P_{y_1}} - P_{y_1}) + \left.\frac{\partial \widehat{\omega}(\widehat{P_{y_1}}, \widehat{P_{y_2}})}{\partial \widehat{P_{y_2}}}\right|_{P_{y_1}, P_{y_2}} (\widehat{P_{y_2}} - P_{y_2}). \tag{3.10}$$

Thus we have

$$\begin{aligned}\widehat{\omega}_{gu} \approx & \frac{P_{y_1} - P_{y_2}}{(p_2 - p_1)\pi_a + (1-p_2)P_{y_1} - (1-p_1)P_{y_2}} + \frac{(p_2 - p_1)(\pi_a - P_{y_2})(\widehat{P_{y_1}} - P_{y_1})}{\{(p_2 - p_1)\pi_a + (1-p_2)P_{y_1} - (1-p_1)P_{y_2}\}^2} \\ & - \frac{(p_2 - p_1)(\pi_a - P_{y_1})(\widehat{P_{y_2}} - P_{y_2})}{\{(p_2 - p_1)\pi_a + (1-p_2)P_{y_1} - (1-p_1)P_{y_2}\}^2}.\end{aligned} \tag{3.11}$$

In Gupta, Tuck, Spears Gill, and Crowe (2013), it is also shown that up to first-order Taylor approximation, $\widehat{\omega}_{gu}$ is an unbiased estimator for $\omega$,

$$E\left(\widehat{\omega}_{gu}\right) \approx \frac{P_{y_1} - P_{y_2}}{(p_2 - p_1)\pi_a + (1-p_2)P_{y_1} - (1-p_1)P_{y_2}} \quad \left(\because E\left[\widehat{P_{y_i}} - P_{y_i}\right] = 0\right). \tag{3.12}$$

Also, up to first-order Taylor approximation, its variance is given by

$$Var\left(\widehat{\omega}_{gu}\right) \approx \frac{(p_2 - p_1)^2 \left\{(\pi_a - P_{y_2})^2 \frac{P_{y_1}(1-P_{y_1})}{n_1} + (\pi_a - P_{y_1})^2 \frac{P_{y_2}(1-P_{y_2})}{n_2}\right\}}{\{(p_2 - p_1)\pi_a + (1-p_2)P_{y_1} - (1-p_1)P_{y_2}\}^4}. \tag{3.13}$$

Notice that the optimal sample size ratio of the optional unrelated RRT model for binary response is given by $\frac{n_2}{n_1} = \lambda\sqrt{\frac{P_{y_2}(1-P_{y_2})}{P_{y_1}(1-P_{y_1})}}$, which generates the smallest possible value of $Var(\widehat{\pi}_{gu})$.

### 3.3 Two-Stage Optional Warner's RRT

This RRT model of Sihm and Gupta (2015) is a mixture of three RRT methods—the indirect response technique of Warner (1965), the two-stage RRT model of Mangat and Singh (1990), and the optional RRT model of Gupta (2001). It contains two randomization devices. The first device is the one used in the two-stage RRT model of Mangat and Singh (1990) and the second is the one used in the optional RRT model of Gupta (2001).

Let $T$ be the pre-assigned probability of asking the sensitive characteristic of the respondents directly in the first stage. Also let $\pi$ be the unknown prevalence of the sensitive characteristic in the population, $p$ be the pre-assigned probability of the respondent selecting the sensitive question directly in the second stage, and $\omega$ be the unknown sensitivity level of the survey question in the population.

The probability of "Yes" response ($P_y$) in this model can be expressed as

$$P_y = T\pi + (1-T)\left\{(1-\omega)\pi + \omega\{p\pi + (1-p)(1-\pi)\}\right\}. \tag{3.14}$$

Using two independent samples with sizes $n_1$ and $n_2$ respectively, and assuming that $p_1$ and $p_2$ are two different pre-assigned probabilities of the respondents selecting the sensitive question in the two samples, Equation (3.14) can be written as

$$P_{y_1} - \pi = (1-T)(p_1-1)(2\pi-1)\omega \qquad \text{and} \qquad P_{y_2} - \pi = (1-T)(p_2-1)(2\pi-1)\omega. \tag{3.15}$$

Solving for $\pi$, we have

$$\pi = \frac{\lambda P_{y_2} - P_{y_1}}{\lambda - 1} \quad \left(p_1 \neq 1, p_2 \neq 1, p_1 \neq p_2, \pi \neq \frac{1}{2}, \text{ and } T \neq 1\right), \quad \text{where } \lambda = \frac{(1-p_1)}{(1-p_2)}. \quad (3.16)$$

Equation (3.16) leads to the following unbiased estimator for $\pi$, given by

$$\widehat{\pi} = \frac{\lambda \widehat{P_{y_2}} - \widehat{P_{y_1}}}{\lambda - 1}, \quad (3.17)$$

with its variance given by

$$Var(\widehat{\pi}) = \frac{1}{(\lambda-1)^2}\left\{\lambda^2 \frac{P_{y_2}(1-P_{y_2})}{n_2} + \frac{P_{y_1}(1-P_{y_1})}{n_1}\right\}. \quad (3.18)$$

The optimal sample sizes $n_1$ and $n_2$ for the two independent samples to minimize $Var\left(\widehat{\pi}\right)$ are given by

$$\left(\frac{n_1}{n_2}\right)_{opt(\widehat{\pi})} = \frac{1}{\lambda}\sqrt{\frac{P_{y_1}(1-P_{y_1})}{P_{y_2}(1-P_{y_2})}}. \quad (3.19)$$

Similarly from Equations (3.15), we have

$$\omega = \frac{P_{y_1} - P_{y_2}}{(1-T)\left\{(1-p_2)(2P_{y_1}-1) - (1-p_1)(2P_{y_2}-1)\right\}} \quad \left(p_1 \neq p_2, T \neq 1, \text{ and } \pi \neq \frac{1}{2}\right), \quad (3.20)$$

which leads to an estimator of $\omega$ given by

$$\widehat{\omega} = \frac{\widehat{P_{y_1}} - \widehat{P_{y_2}}}{(1-T)\left\{(1-p_2)(2\widehat{P_{y_1}}-1) - (1-p_1)(2\widehat{P_{y_2}}-1)\right\}}. \tag{3.21}$$

Sihm and Gupta (2015) approximate the mean and variance of this estimator by using first-order Taylor expansion. Let us rewrite $\widehat{\omega}$ as

$$\widehat{\omega} \approx \widehat{\omega}(P_{y_1}, P_{y_2}) + \left.\frac{\partial \widehat{\omega}(\widehat{P_{y_1}}, \widehat{P_{y_2}})}{\partial \widehat{P_{y_1}}}\right|_{P_{y_1}, P_{y_2}} (\widehat{P_{y_1}} - P_{y_1}) + \left.\frac{\partial \widehat{\omega}(\widehat{P_{y_1}}, \widehat{P_{y_2}})}{\partial \widehat{P_{y_2}}}\right|_{P_{y_1}, P_{y_2}} (\widehat{P_{y_2}} - P_{y_2}). \tag{3.22}$$

With $(\theta_i = 2P_{y_i} - 1)$, we have

$$\widehat{\omega} \approx \frac{1}{2(1-T)}\left\{\frac{\theta_1-\theta_2}{(1-p_2)\theta_1-(1-p_1)\theta_2} + \frac{(p_1-p_2)\theta_2(\widehat{\theta_1}-\theta_1)}{\{(1-p_2)\theta_1-(1-p_1)\theta_2\}^2} - \frac{(p_1-p_2)\theta_1(\widehat{\theta_2}-\theta_2)}{\{(1-p_2)\theta_1-(1-p_1)\theta_2\}^2}\right\}. \tag{3.23}$$

In Sihm and Gupta (2015), it is also shown that up to first-order Taylor approximation, $\widehat{\omega}$ is an unbiased estimator for $\omega$,

$$E\left(\widehat{\omega}\right) \approx \frac{P_{y_1} - P_{y_2}}{(1-T)\{(1-p_2)\theta_1 - (1-p_1)\theta_2\}} \qquad \left(\because E\left[\widehat{\theta_i} - \theta_i\right] = 0\right). \tag{3.24}$$

Also, up to first-order Taylor approximation, its variance is given by

$$Var\left(\widehat{\omega}\right) \approx \frac{(p_1-p_2)^2\left\{(2P_{y_2}-1)^2\frac{P_{y_1}(1-P_{y_1})}{n_1} + (2P_{y_1}-1)^2\frac{P_{y_2}(1-P_{y_2})}{n_2}\right\}}{(1-T)^2\{(1-p_2)(2P_{y_1}-1)-(1-p_1)(2P_{y_2}-1)\}^4}. \tag{3.25}$$

Notice that the optimal sample size ratio of the two-stage optional Warner's RRT model for binary response is given by $\frac{n_2}{n_1} = \lambda\sqrt{\frac{P_{y_2}(1-P_{y_2})}{P_{y_1}(1-P_{y_1})}}$, which generates the smallest possible value of $Var(\widehat{\pi})$.

Sihm and Gupta (2015) proved that the variance of the estimator $\widehat{\pi}$ from Equation (3.18) in this section can be always made smaller than the variance of the estimator $\widehat{\pi}_g$ from Equation (2.50) of Gupta (2001) in Section 2.4.1.

Using Equations (2.50) and (3.18), we can solve the inequality such that

$$Var(\widehat{\pi}) < Var\left(\widehat{\pi}_g\right)$$

$$\Longrightarrow \quad \frac{n_1\lambda\big\{2(1-p_2)\omega - 1\big\} + n_2\big\{2(1-p_1)\omega - 1\big\}}{\omega n(1-p_1)} < T < 1 \tag{3.26}$$

$$\left(\text{with} \quad \lambda = \frac{p_1 - 1}{p_2 - 1}\right).$$

Thus, a meaningful value of $T$ that satisfies Equation (3.26) can always be chosen (Sihm and Gupta, 2015).

### 3.4 Efficiency Comparisons

For all the tables in Chapter V (Simulation Results), the first two columns are for Section 3.2 RRT model of Gupta, Tuck, Spears Gill, and Crowe (2013) and Section 3.3 RRT model of Sihm and Gupta (2015). The variances of the estimators from these two models are not much different, and most of the time, the Section 3.2 binary optional unrelated RRT model of Gupta, Tuck, Spears Gill, and Crowe (2013) performs a little better than the Section 3.3 model, with a small value of $T$; however this can always be reversed by assigning a bigger value of $T$ by the interviewer.

For Tables 1 - 4 ($\pi = 0.1, \omega = 0.1, 0.3, 0.7, 0.9$), simulated values and theoretical values of $Var(\pi_{gu})$ of Gupta, Tuck, Spears Gill, and Crowe (2013) in Section 3.2 and those of $Var(\pi)$ of Sihm and Gupta (2015) in Section 3.3 are very similar to each other. The similarity between $Var(\pi_{gu})$ and $Var(\pi)$ doesn't change for Tables 5 - 12 either; but, there are a few things to notice in Tables 5 - 12. For Tables 5 - 8 ($\pi = 0.3, \omega = 0.1, 0.3, 0.7, 0.9$), simulated values of $E(\widehat{\omega})$ are sometimes negative or even greater than 1 while $Var(\omega)$ gets very volatile for the model of Gupta, Tuck, Spears Gill, and Crowe (2013) in Section 3.2. This is because the combination of pre-assigned proportions as well as simulated parameter values make the denominator of Equation (3.8) close to zero, increasing variability of the estimator for $\omega$. Likewise for Tables 9 - 12 ($\pi = 0.6, \omega = 0.1, 0.3, 0.7, 0.9$), the same thing occurs to the model of Sihm and Gupta (2015) in Section 3.3, making the denominator of Equation (3.20) close to zero. Thus, we have negative estimates for the proportion as well as very big values of variance of the estimator $\widehat{\omega}$.

By carefully assigning $p_1$ and $p_2$ so that Equations (3.8) and (3.20) do not have division by zero, we can avoid these unfortunate cases in which the amount of variance spikes due to small or near zero value of the denominator. Other than that, the two models covered in this chapter deliver very similar performance when it comes to estimating the prevalence of sensitive characteristic and the level of sensitivity.

## 3.5 Conclusion

As one can easily notice from Tables 1 - 12 in Chapter V, the RRT methods in this chapter do not outperform other RRT methods in Chapter IV. By splitting the given sample, the methods in this chapter reduce the amount of available sample in order to estimate two parameters simultaneously. In the next chapter, we discuss a new technique that does not increase the sampling burden while estimating two parameters simultaneously.

# CHAPTER IV
# MODIFIED BINARY OPTIONAL RRT MODELS WITH TWO-QUESTION APPROACH

## 4.1 Introduction

Our proposed model is a modified two-stage binary optional RRT model based on the RRT model of Sihm and Gupta (2015). Unlike the model of Sihm and Gupta (2015), which uses the split-sample approach to estimate $\pi$ and $\omega$, the new model asks two separate questions of the same sample of respondents. Question 1 is the auxiliary question to estimate the level of sensitivity of the main research question in the population while Question 2 is the main research question to estimate the prevalence of the sensitive characteristic in the population. This technique was first explored in Sihm, Chhabra, and Gupta (2016) and allowed a smaller sample size for a given efficiency level.

## 4.2 Binary Optional Unrelated RRT with Two-Question Approach

The main motivation for the binary model of Sihm, Chhabra, and Gupta (2016) is to avoid the split-sample approach, which requires a larger sample size. This is done by asking respondents two separate questions. Question 1 is the auxiliary question about whether or not the main research question is sensitive enough for the respondent to opt for a scrambled response. Question 2 is the main research question which the respondent answers by using the optional binary unrelated question RRT model of Gupta, Tuck, Spears Gill, and Crowe (2013) in Section 3.2. Respondents will answer Question 1 by using the original binary unrelated question RRT model of Greenberg et al. (1969).

Let $\pi_a$ be the known prevalence of an unrelated innocuous characteristic, $\pi_b$ be the known prevalence of another unrelated innocuous characteristic, $\pi$ be the unknown prevalence of the sensitive characteristic, $p_a$ be the pre-assigned probability of the respondent selecting the sensitive question in answering Question 1, $p_b$ be the pre-assigned probability of the respondent selecting the question about sensitivity in answering Question 2, and $\omega$ be the unknown sensitivity level of the survey question in the population.

Let $P_{y_i}$ be the probability of "Yes" response from a respondent to Question $i$ $(i = 1, 2)$. Then,

$$P_{y_1} = p_a\omega + (1 - p_a)\pi_a \tag{4.1}$$

$$P_{y_2} = (1 - \omega)\pi + \omega\{\pi p_b + (1 - p_b)\pi_b\}. \tag{4.2}$$

Solving Equations (4.1) and (4.2) for $\pi$ and $\omega$ respectively, we have

$$\pi = \frac{P_{y_2} - (1 - p_b)\omega\pi_b}{1 - (1 - p_b)\omega}, \qquad \text{and} \qquad \omega = \frac{P_{y_1} - (1 - p_a)\pi_a}{p_a}, \tag{4.3}$$

which lead to the estimators

$$\widehat{\pi} = \frac{\widehat{P_{y_2}} - (1 - p_b)\widehat{\omega}\pi_b}{1 - (1 - p_b)\widehat{\omega}}, \qquad \text{and} \qquad \widehat{\omega} = \frac{\widehat{P_{y_1}} - (1 - p_a)\pi_a}{p_a}, \tag{4.4}$$

where $\widehat{P_{y_i}}$ is the proportion of "Yes" responses in the sample to Question $i$ $(i = 1, 2)$. It is easy to see that $\widehat{\omega}$ is an unbiased estimator:

$$E\left(\widehat{\omega}\right) = E\left(\frac{\widehat{P_{y_1}} - (1-p_a)\pi_a}{p_a}\right) = \frac{E\left(\widehat{P_{y_1}}\right) - (1-p_a)\pi_a}{p_a} = \frac{P_{y_1} - (1-p_a)\pi_a}{p_a} = \omega. \tag{4.5}$$

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

$$Var\left(\widehat{\omega}\right) = Var\left(\frac{\widehat{P_{y_1}} - (1-p_a)\pi_a}{p_a}\right) = \frac{Var\left(\widehat{P_{y_1}}\right) - (1-p_a)\pi_a}{p_a} = \frac{P_{y_1}(1-P_{y_1})}{np_a^2}. \tag{4.6}$$

Sihm, Chhabra, and Gupta (2016) approximate the mean and variance of the estimator for $\pi$ by using first-order Taylor expansion. We first rewrite $\pi$ as

$$\widehat{\pi} \quad \approx \quad \widehat{\pi}(P_{y_2}, \omega) + \left.\frac{\partial\widehat{\pi}(\widehat{P_{y_2}}, \widehat{\omega})}{\partial\widehat{P_{y_2}}}\right|_{P_{y_2},\omega} (\widehat{P_{y_2}} - P_{y_2}) + \left.\frac{\partial\widehat{\pi}(\widehat{P_{y_2}}, \widehat{\omega})}{\partial\widehat{\omega}}\right|_{P_{y_2},\omega} (\widehat{\omega} - \omega). \tag{4.7}$$

After getting the partial derivatives, we have

$$\widehat{\pi} \quad \approx \quad \frac{P_{y_2} - \omega(1-p_b)\pi_b}{1-(1-p_b)\omega} + \frac{\widehat{P_{y_2}} - P_{y_2}}{1-(1-p_b)\omega} + \frac{(1-p_b)(P_{y_2} - \pi_b)(\widehat{\omega} - \omega)}{\{1-(1-p_b)\omega\}^2}. \tag{4.8}$$

Sihm, Chhabra, and Gupta (2016) show that up to first-order Taylor approximation, $\widehat{\pi}$ is an unbiased estimator for $\pi$,

$$E\left(\pi\right) \quad \approx \quad \frac{P_{y_2} - \omega(1-p_b)\pi_b}{1-(1-p_b)\omega} \qquad \Big(\because \quad E\left[\widehat{P_{y_2}} - P_{y_2}\right] = 0 \quad \text{and} \quad E\left[\widehat{\omega} - \omega\right] = 0\Big). \tag{4.9}$$

Also, up to first-order Taylor approximation, its variance is given by

$$Var\,(\widehat{\pi}) \approx \frac{1}{\{1-(1-p_b)\omega\}^2}\left\{\frac{P_{y_2}(1-P_{y_2})}{n}\right\} + \frac{(1-p_b)^2(P_{y_2}-\pi_b)^2}{\{1-(1-p_b)\omega\}^4}\left\{\frac{P_{y_1}(1-P_{y_1})}{np_a^2}\right\}. \tag{4.10}$$

## 4.3 Revised Binary Optional Unrelated RRT with Two-Question Approach

For the model discussed in the previous section, we assume that $\pi_a$ and $\pi_b$ are somehow known, but it is more realistic to assume that they are unknown when the survey is being done. If both $\pi_a$ and $\pi_b$ are assumed to be unknown, we can still estimate $\pi$ and $\omega$ by using a combination of the split-sample and two-question approaches.

The probability of "Yes" response $(P_y)$ to the auxiliary question in this model can be expressed as

$$P_{y_1} = p_a\omega + (1-p_a)\pi_a. \tag{4.11}$$

Using two independent samples with sizes $n_1$ and $n_2$ respectively, and assuming that $p_{a_1}$ and $p_{a_2}$ are two different pre-assigned probabilities of the respondents selecting the sensitive question in the two samples, Equation (4.11) can be written as

$$P_{y_{11}} = p_{a_1}\omega + (1-p_{a_1})\pi_a \qquad \text{and} \qquad P_{y_{12}} = p_{a_2}\omega + (1-p_{a_2})\pi_a. \tag{4.12}$$

Solving for $\omega$, we have

$$\omega = \frac{\lambda_a P_{y_{12}} - P_{y_{11}}}{\lambda_a p_{a_2} - p_{a_1}} \quad (p_{a_1} \neq 1, p_{a_2} \neq 1, \text{ and } p_{a_1} \neq p_{a_2}), \quad \text{where } \lambda_a = \frac{(1 - p_{a_1})}{(1 - p_{a_2})}. \tag{4.13}$$

Equation (4.13) leads to the following unbiased estimator for $\omega$, given by

$$\widehat{\omega} = \frac{\lambda_a \widehat{P_{y_{12}}} - \widehat{P_{y_{11}}}}{\lambda_a p_{a_2} - p_{a_1}}, \tag{4.14}$$

with its variance given by

$$Var(\widehat{\omega}) = \frac{1}{(\lambda_a p_{a_2} - p_{a_1})^2} \left\{ \lambda_a^2 \frac{P_{y_{12}}(1 - P_{y_{12}})}{n_2} + \frac{P_{y_{11}}(1 - P_{y_{11}})}{n_1} \right\}. \tag{4.15}$$

The probability of "Yes" response $(P_y)$ to the main research question can be expressed as

$$P_{y_2} = (1 - \omega)\pi + \omega\{p\pi + (1 - p)\pi_b\}. \tag{4.16}$$

Using the same two samples, and assuming that $p_{b_1}$ and $p_{b_2}$ are two different pre-assigned probabilities of the respondents selecting the sensitive question in the two samples, Equation (4.16) can be written as

$$P_{y_{21}} = \pi + \omega(1 - p_{b_1})(\pi_b - \pi) \qquad \text{and} \qquad P_{y_{22}} = \pi + \omega(1 - p_{b_2})(\pi_b - \pi). \tag{4.17}$$

Solving for $\pi$, we have

$$\pi = \frac{\lambda_b P_{y_{22}} - P_{y_{21}}}{\lambda_b - 1} \quad (p_{b_1} \neq 1, p_{b_2} \neq 1, p_{b_1} \neq p_{b_2}, \ \text{ and } \ \pi_b \neq \pi, ), \quad \text{ where } \lambda_b = \frac{(1-p_{b_1})}{(1-p_{b_2})}. \tag{4.18}$$

This leads to the following unbiased estimator for $\pi$, given by

$$\widehat{\pi}_r = \frac{\lambda_b \widehat{P_{y_{22}}} - \widehat{P_{y_{21}}}}{\lambda_b - 1}, \tag{4.19}$$

with its variance given by

$$Var(\widehat{\pi}_r) = \frac{1}{(\lambda_b - 1)^2} \left\{ \lambda_b^2 \frac{P_{y_{22}}(1-P_{y_{22}})}{n_2} + \frac{P_{y_{21}}(1-P_{y_{21}})}{n_1} \right\}. \tag{4.20}$$

The optimal sample sizes $n_1$ and $n_2$ are used for the two independent samples to minimize $Var\,(\widehat{\pi}_r)$ and the ratio is given by

$$\left(\frac{n_1}{n_2}\right)_{opt(\widehat{\pi}_r)} = \frac{1}{\lambda_b} \sqrt{\frac{P_{y_{21}}(1-P_{y_{21}})}{P_{y_{22}}(1-P_{y_{22}})}}. \tag{4.21}$$

## 4.4 Two-Stage Optional Indirect RRT with Two-Question Approach

The underlying structure of our proposed model in this section is the same as in Sihm and Gupta (2015) model of Section 3.3, except that we now employ the two-question technique of Sihm, Chhabra, and Gupta (2016) from Section 4.2 instead of using the split-sample approach.

Here, we ask two separate questions of the same sample. With Question 1, we estimate the level of sensitivity to the main research question among the population by using the indirect question RRT model of Warner (1965). With Question 2, we estimate the prevalence of the sensitive characteristic associated with the main research question by using the same two-stage binary optional indirect question RRT model as in Sihm and Gupta (2015), which was discussed in Section 3.3.

Let $p_a$ be the pre-assigned probability of the respondent selecting the direct question about sensitivity in answering Question 1, $\omega$ be the unknown sensitivity level of the survey question in the population, $T$ be the pre-assigned probability of asking the sensitive characteristic of the respondents directly in the first stage of answering Question 2, $\pi$ be the unknown prevalence of the sensitive characteristic, and $p_b$ be the pre-assigned probability of the respondent selecting the sensitive question in the second stage of answering Question 2.

Let $P_{y_i}$ be the probability of "Yes" response from a respondent to Question $i$ $(i = 1, 2)$, we have

$$P_{y_1} = p_a\omega + (1 - p_a)(1 - \omega) \tag{4.22}$$

$$P_{y_2} = T\pi + (1 - T)\left\{(1 - \omega)\pi + \omega\{p_b\pi + (1 - p_b)(1 - \pi)\}\right\}. \tag{4.23}$$

Solving Equations (4.22) and (4.23) for $\pi$ and $\omega$ respectively, we have

$$\omega = \frac{P_{y_1} - (1 - p_a)}{2p_a - 1}, \qquad \text{and} \qquad \pi = \frac{P_{y_2} - (1 - T)(1 - p_b)\omega}{1 - 2(1 - T)(1 - p_b)\omega} \qquad \left(p_a \neq \frac{1}{2}\right), \tag{4.24}$$

which lead to the estimators

$$\widehat{\omega} = \frac{\widehat{P_{y_1}} - (1-p_a)}{2p_a - 1} \qquad \text{and} \qquad \widehat{\pi}_p = \frac{\widehat{P_{y_2}} - (1-T)(1-p_b)\widehat{\omega}}{1 - 2(1-T)(1-p_b)\widehat{\omega}} \qquad \left(p_a \neq \frac{1}{2}\right), \tag{4.25}$$

where $\widehat{P_{y_i}}$ is the proportion of "Yes" responses in the sample to Question $i$ $(i = 1, 2)$. Notice that $\widehat{\omega}$ is an unbiased estimator for $\omega$ such that

$$E\left(\widehat{\omega}\right) = E\left(\frac{\widehat{P_{y_1}} - (1-p_a)}{2p_a - 1}\right) = \frac{E\left(\widehat{P_{y_1}}\right) - (1-p_a)}{2p_a - 1} = \frac{P_{y_1} - (1-p_a)}{2p_a - 1} = \omega. \tag{4.26}$$

Its variance is given by

$$Var\left(\widehat{\omega}\right) = Var\left(\frac{\widehat{P_{y_1}} - (1-p_a)}{2p_a - 1}\right) = \frac{P_{y_1}(1-P_{y_1})}{n(2p_a - 1)^2}. \tag{4.27}$$

The mean and variance of the estimator for $\pi$ is estimated by first-order Taylor expansion. We first rewrite $\widehat{\pi}_p$ as

$$\widehat{\pi}_p \quad \approx \quad \widehat{\pi}(P_{y_2}, \omega) + \left.\frac{\partial \widehat{\pi}(\widehat{P_{y_2}}, \widehat{\omega})}{\partial \widehat{P_{y_2}}}\right|_{P_{y_2}, \omega} (\widehat{P_{y_2}} - P_{y_2}) + \left.\frac{\partial \widehat{\pi}(\widehat{P_{y_2}}, \widehat{\omega})}{\partial \widehat{\omega}}\right|_{P_{y_2}, \omega} (\widehat{\omega} - \omega). \tag{4.28}$$

Thus, we have

$$\widehat{\pi}_p \approx \frac{P_{y_2} - (1-T)(1-p_b)\omega}{1 - 2(1-T)(1-p_b)\omega} + \frac{\widehat{P_{y_2}} - P_{y_2}}{1 - 2(1-T)(1-p_b)\omega} + \frac{(1-T)(1-p_b)(2P_{y_2} - 1)(\widehat{\omega} - \omega)}{\{1 - 2(1-T)(1-p_b)\omega\}^2}. \tag{4.29}$$

Up to first-order Taylor approximation, $\widehat{\pi}_p$ is an unbiased estimator for $\pi$,

$$E\left(\widehat{\pi}_p\right) \approx \frac{P_{y_2} - (1-T)(1-p_b)\omega}{1-2(1-T)(1-p_b)\omega} \qquad \left(\because \quad E\left(\widehat{P_{y_2}} - P_{y_2}\right) = 0 \quad \text{and} \quad E\left(\widehat{\omega} - \omega\right) = 0\right). \tag{4.30}$$

Also, up to first-order Taylor expansion, its variance is given by

$$Var\left(\widehat{\pi}_p\right) \approx \frac{P_{y_2}(1-P_{y_2})}{n\{1-2(1-T)(1-p_b)\omega\}^2} + \frac{(1-T)^2(1-p_b)^2(2P_{y_2}-1)^2 P_{y_1}(1-P_{y_1})}{n\{1-2(1-T)(1-p_b)\omega\}^4(2p_a-1)^2}. \tag{4.31}$$

## 4.5 Efficiency Comparisons

From Tables 1 - 12 in Chapter V, one can easily see that the proposed model of Section 4.4 performs as well as the binary optional unrelated question RRT with two questions in Section 4.3, except for occasional spikes of variances of the estimators due to near zero value of the denominator of the estimator caused by unfortunate combination of pre–assigned proportions. This can be easily avoided in real life scenarios, simply by plugging in those values to the formula such as Equation (4.25) to see if the denominator gets closer to zero or not.

Now let us develop a more systematical way of determining suitable interval of $T$ in contrast with other parameters to get a more efficient estimator. Let us compare the theoretical variance of each estimator for $\pi$ from Section 4.3 and Section 4.4. For the proposed model, we have the variance of the estimator $\widehat{\pi}_p$ from Equation (4.31),

$$Var\left(\widehat{\pi}_p\right) \approx \frac{P_{y_2}(1-P_{y_2})}{n\{1-2(1-T)(1-p_b)\omega\}^2} + \frac{(1-T)^2(1-p_b)^2(2P_{y_2}-1)^2 P_{y_1}(1-P_{y_1})}{n\{1-2(1-T)(1-p_b)\omega\}^4(2p_a-1)^2}. \tag{4.32}$$

For the revised model of Sihm, Chhabra, and Gupta (2016) , we have the variance of the estimator $\widehat{\pi}_r$ from Equation (4.20),

$$Var(\widehat{\pi}_r) = \frac{1}{(\lambda_b - 1)^2} \left\{ \lambda_b^2 \frac{P_{y_{22}}(1 - P_{y_{22}})}{n_2} + \frac{P_{y_{21}}(1 - P_{y_{21}})}{n_1} \right\}. \tag{4.33}$$

Solving the inequality $Var\,(\widehat{\pi}_p) < Var(\widehat{\pi}_r)$ with respect to $T$ would be quite a complex undertaking. Sihm and Gupta (2015) actually did very similar thing to get a suitable interval of $T$ and eventually proved that it's always possible to get an appropriate $T$ to have smaller variance than other models. But it was a quadratic equation in $T$. Here, we have a quartic expression in $T$ to solve the inequality. Instead of solving it and get an algebraic solution for the suitable interval of $T$ which guarantees smaller variance of the estimator, let us again use computer simulation to figure out where the interval of $T$ will reside, especially in combination with another parameter $p_a$ to which we can assign any value.

The R code for this efficiency comparison is listed in APPENDIX B. Let us first run a ‘‘hit or miss’’ simulation on Table 1. For every value of $T$, Table 1 shows that $Var\,(\widehat{\pi}_p)$ is always smaller than $Var(\widehat{\pi}_r)$. Thus, we don't have to actively seek an appropriate value of $T$ to get a smaller variance on Table 1. Figure 1 clearly shows this. The blue line in Figure 1 indicates the pre-assigned value of $p_a = 0.8$. The red area is where $Var\,(\widehat{\pi}_p)$ is smaller than $Var(\widehat{\pi}_r)$ and for the black area, it's the opposite. From Figure 1, it's clear that every value of $T$ will lead to $Var\,(\widehat{\pi}_p) < Var(\widehat{\pi}_r)$ as $p_a = 0.8$.

On the other hand, Tables 4, 8, and 12 show the opposite cases. These three cases all have $\omega = 0.9$, which means the level of sensitivity is extremely high and almost all of the respondents prefer using randomization device. In these cases, we can still select meaningful value of $T$, for example $T = 0.6$, to achieve better efficiency with the proposed model. Figures 4, 8, and 12 display this graphically.

### 4.6 Conclusion

In this chapter, we first presented the model of Sihm, Chhabra, and Gupta (2016) in Section 4.2 to show how much improvement in efficiency we could get by switching from the split-sample approach to the two-question approach. But the model of Sihm, Chhabra, and Gupta (2016) has one drawback that requires the researcher to know the prevalence of the innocuous characters ($\pi_a$ and $\pi_b$) among the population before carrying out survey study. Instead of assuming $\pi_a$ and $\pi_b$ to be known, we discussed a more realistic revision of the model in Section 4.3. After that, we proposed our final binary optional RRT model in Section 4.4.

The nice thing about our proposed model in Section 4.4 is that it doesn't require the demanding assumption of exactly knowing the prevalence of an innocuous characteristic unlike most of the models associated with unrelated question models. We showed that by simply asking one more question, we are able to significantly improve the efficiency of our estimator. In addition to that, we propose a new method that doesn't require the onerous condition of knowing the other parameters to properly use the method. Using simulation study, as shown in Section 5.3, we demonstrated that we can always select a value of $T$ which will improve the efficiency of the new estimator as compared to the revised model of Sihm, Chhabra, and Gupta (2016) in Section 4.3.

# CHAPTER V

# SIMULATION RESULTS

## 5.1 Introduction

In this simulation study, we included a more realistic version of Sihm, Chhabra, and Gupta (2016) model in the second column from the right. This model was discussed in detail in Section 4.3 and assumes the prevalence of innocuous characteristics $\pi_a$ and $\pi_b$ are actually unknown to us. We assigned the following values to the various parameters: $p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$. The number of trials for each table is set to be 10,000. Statistical software `R version 3.3.3` (R Core Team, 2017) was used for all the simulations in this study.

These parameters can be assigned at our will and can be changed without difficulties. Notice that we can also set the value of the two-stage parameter $T$ freely before any survey begins.

Due to the algebraic limitations (division by zero), from time to time, some combinations of these parameters produce unusually high variances. These incidents are entirely avoidable because, in the first place, we can clearly measure how closely the denominator approaches zero by substituting these values into the estimators. For examples, let us talk about Equation (4.24).

$$\omega = \frac{P_{y_1} - (1-p_a)}{2p_a - 1}, \qquad \text{and} \qquad \pi = \frac{P_{y_2} - (1-T)(1-p_b)\omega}{1 - 2(1-T)(1-p_b)\omega} \qquad \left(p_a \neq \frac{1}{2}\right)$$

If we assign a value close to 0.5 to $p_a$, the estimator for $\omega$ will be extremely unstable and will produce a large variance. Likewise if we happen to have all the parameter values in such a way that $(1-2(1-T)(1-p_b)\omega)$ gets very close to zero, then the estimator would be useless. That's exactly what happened to Tables 4, 7, 8, 11, and 12. In Table 4, when $T = 0.2$, the theoretical variance and empirical variance of the proposed model suddenly became large. If all the parameter values are plugged in to the denominator, it will be $1-2\times(1-0.2)(1-0.3)\times 0.9 = -0.008$, which is very close to zero. And this creates unusually high values of the empirical and theoretical variances for the estimator.

## 5.2 Simulation Results to Compare Efficiency

For Tables 1 - 12, the first column is for the model of Gupta, Tuck, Spears Gill, and Crowe (2013) in Section 3.2 and the second for the model of Sihm and Gupta (2015) in Section 3.3. These two models are based on the split-sample approach and the simulation results clearly show their bigger variances of the estimators. The third column from the left is for the model of Sihm, Chhabra, and Gupta (2016) in Section 4.2; the fourth for the revised model of Sihm, Chhabra, and Gupta (2016) in Section 4.3 with realistic assumptions of unknown $\pi_a$ and $\pi_b$; the fifth for the propose model in Section 4.4. Notice that even though the model of Sihm, Chhabra, and Gupta (2016) originally didn't use the split-sample approach, the revised model has to use it as there are four unknown parameters in the model and that's the only way to estimate them all.

For those models with the split-sample approach, the optimal ratios of $n_1/n_2$ were calculated and two sub-samples were grouped that way.

Table 1. Comparison of Different Two-Stage Model Estimators ($\pi = 0.1$ & $\omega = 0.1$)

| $\pi = 0.1,\ \omega = 0.1$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)[*] | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.099924 | 0.100342 | 0.099952 | 0.099975 | 0.099766 |
| Empirical $Var(\widehat{\pi})$ | 0.000314 | 0.000356 | 0.000112 | 0.000866 | 0.000330 |
| Theoretical $Var(\widehat{\pi})$ | 0.000311 | 0.000352 | 0.000116 | 0.000853 | 0.000324 |
| Empirical $Mean(\widehat{\omega})$ | 0.089509 | 0.096518 | 0.099871 | 0.099927 | 0.099512 |
| Empirical $Var(\widehat{\omega})$ | 0.035684 | 0.003789 | 0.000205 | 0.000412 | 0.000530 |
| Theoretical $Var(\widehat{\omega})$ | 0.034356 | 0.003723 | 0.000199 | 0.000411 | 0.000534 |
| Optimal $n_1$ | 769 | 755 | | 663 | |
| Optimal $n_2$ | 231 | 245 | | 337 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.099924 | 0.099996 | 0.099952 | 0.099975 | 0.100065 |
| Empirical $Var(\widehat{\pi})$ | 0.000314 | 0.000341 | 0.000112 | 0.000866 | 0.000249 |
| Theoretical $Var(\widehat{\pi})$ | 0.000311 | 0.000341 | 0.000116 | 0.000853 | 0.000253 |
| Empirical $Mean(\widehat{\omega})$ | 0.089509 | 0.097169 | 0.099871 | 0.099927 | 0.099233 |
| Empirical $Var(\widehat{\omega})$ | 0.035684 | 0.005695 | 0.000205 | 0.000412 | 0.000537 |
| Theoretical $Var(\widehat{\omega})$ | 0.034356 | 0.005717 | 0.000199 | 0.000411 | 0.000534 |
| Optimal $n_1$ | 769 | 759 | | 663 | |
| Optimal $n_2$ | 231 | 241 | | 337 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.099924 | 0.100218 | 0.099952 | 0.099975 | 0.099740 |
| Empirical $Var(\widehat{\pi})$ | 0.000314 | 0.000330 | 0.000112 | 0.000866 | 0.000192 |
| Theoretical $Var(\widehat{\pi})$ | 0.000311 | 0.000329 | 0.000116 | 0.000853 | 0.000194 |
| Empirical $Mean(\widehat{\omega})$ | 0.089509 | 0.095074 | 0.099871 | 0.099927 | 0.100354 |
| Empirical $Var(\widehat{\omega})$ | 0.035684 | 0.010186 | 0.000205 | 0.000412 | 0.000544 |
| Theoretical $Var(\widehat{\omega})$ | 0.034356 | 0.009940 | 0.000199 | 0.000411 | 0.000534 |
| Optimal $n_1$ | 769 | 763 | | 663 | |
| Optimal $n_2$ | 231 | 237 | | 337 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.099924 | 0.100200 | 0.099952 | 0.099975 | 0.100191 |
| Empirical $Var(\widehat{\pi})$ | 0.000314 | 0.000311 | 0.000112 | 0.000866 | 0.000110 |
| Theoretical $Var(\widehat{\pi})$ | 0.000311 | 0.000304 | 0.000116 | 0.000853 | 0.000111 |
| Empirical $Mean(\widehat{\omega})$ | 0.089509 | 0.087559 | 0.099871 | 0.099927 | 0.099703 |
| Empirical $Var(\widehat{\omega})$ | 0.035684 | 0.085368 | 0.000205 | 0.000412 | 0.000528 |
| Theoretical $Var(\widehat{\omega})$ | 0.034356 | 0.084350 | 0.000199 | 0.000411 | 0.000534 |
| Optimal $n_1$ | 769 | 772 | | 663 | |
| Optimal $n_2$ | 231 | 228 | | 337 | |

[*] More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

Table 2. Comparison of Different Two-Stage Model Estimators ($\pi = 0.1$ & $\omega = 0.3$)

| $\pi = 0.1$, $\omega = 0.3$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)* | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.099802 | 0.100093 | 0.100050 | 0.100133 | 0.098768 |
| Empirical $Var(\widehat{\pi})$ | 0.000347 | 0.000455 | 0.000183 | 0.000939 | 0.000788 |
| Theoretical $Var(\widehat{\pi})$ | 0.000349 | 0.000455 | 0.000189 | 0.000936 | 0.000698 |
| Empirical $Mean(\widehat{\omega})$ | 0.290644 | 0.297225 | 0.300188 | 0.300393 | 0.299833 |
| Empirical $Var(\widehat{\omega})$ | 0.036033 | 0.004003 | 0.000340 | 0.000653 | 0.000639 |
| Theoretical $Var(\widehat{\omega})$ | 0.035331 | 0.00397 | 0.000334 | 0.000652 | 0.000654 |
| Optimal $n_1$ | 756 | 737 | | 656 | |
| Optimal $n_2$ | 244 | 263 | | 344 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.099802 | 0.099852 | 0.100050 | 0.100133 | 0.099314 |
| Empirical $Var(\widehat{\pi})$ | 0.000347 | 0.000431 | 0.000183 | 0.000939 | 0.000522 |
| Theoretical $Var(\widehat{\pi})$ | 0.000349 | 0.000426 | 0.000189 | 0.000936 | 0.000494 |
| Empirical $Mean(\widehat{\omega})$ | 0.290644 | 0.297712 | 0.300188 | 0.300393 | 0.299836 |
| Empirical $Var(\widehat{\omega})$ | 0.036033 | 0.006482 | 0.000340 | 0.000653 | 0.000652 |
| Theoretical $Var(\widehat{\omega})$ | 0.035331 | 0.006202 | 0.000334 | 0.000652 | 0.000654 |
| Optimal $n_1$ | 756 | 740 | | 656 | |
| Optimal $n_2$ | 244 | 260 | | 344 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.099802 | 0.100438 | 0.100050 | 0.100133 | 0.100176 |
| Empirical $Var(\widehat{\pi})$ | 0.000347 | 0.000400 | 0.000183 | 0.000939 | 0.000339 |
| Theoretical $Var(\widehat{\pi})$ | 0.000349 | 0.000396 | 0.000189 | 0.000936 | 0.000343 |
| Empirical $Mean(\widehat{\omega})$ | 0.290644 | 0.293717 | 0.300188 | 0.300393 | 0.299564 |
| Empirical $Var(\widehat{\omega})$ | 0.036033 | 0.011197 | 0.000340 | 0.000653 | 0.000648 |
| Theoretical $Var(\widehat{\omega})$ | 0.035331 | 0.010872 | 0.000334 | 0.000652 | 0.000654 |
| Optimal $n_1$ | 756 | 745 | | 656 | |
| Optimal $n_2$ | 244 | 255 | | 344 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.099802 | 0.100388 | 0.100050 | 0.100133 | 0.100036 |
| Empirical $Var(\widehat{\pi})$ | 0.000347 | 0.000330 | 0.000183 | 0.000939 | 0.000142 |
| Theoretical $Var(\widehat{\pi})$ | 0.000349 | 0.000329 | 0.000189 | 0.000936 | 0.000146 |
| Empirical $Mean(\widehat{\omega})$ | 0.290644 | 0.285063 | 0.300188 | 0.300393 | 0.299861 |
| Empirical $Var(\widehat{\omega})$ | 0.036033 | 0.090200 | 0.000340 | 0.000653 | 0.000650 |
| Theoretical $Var(\widehat{\omega})$ | 0.035331 | 0.089462 | 0.000334 | 0.000652 | 0.000654 |
| Optimal $n_1$ | 756 | 763 | | 656 | |
| Optimal $n_2$ | 244 | 237 | | 344 | |

* More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

## Table 3. Comparison of Different Two-Stage Model Estimators ($\pi = 0.1$ & $\omega = 0.7$)

| $\pi = 0.1$, $\omega = 0.7$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)* | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.100109 | 0.099857 | 0.099671 | 0.099879 | 0.313794 |
| Empirical $Var(\widehat{\pi})$ | 0.000419 | 0.000584 | 0.000553 | 0.001071 | 9.769110 |
| Theoretical $Var(\widehat{\pi})$ | 0.000416 | 0.000596 | 0.000567 | 0.001093 | 0.624840 |
| Empirical $Mean(\widehat{\omega})$ | 0.687067 | 0.698045 | 0.700115 | 0.699954 | 0.700080 |
| Empirical $Var(\widehat{\omega})$ | 0.036029 | 0.003136 | 0.000371 | 0.000702 | 0.000663 |
| Theoretical $Var(\widehat{\omega})$ | 0.034358 | 0.003127 | 0.000364 | 0.000717 | 0.000654 |
| Optimal $n_1$ | 742 | 741 | | 647 | |
| Optimal $n_2$ | 258 | 259 | | 353 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.100109 | 0.099873 | 0.099671 | 0.099879 | 0.094070 |
| Empirical $Var(\widehat{\pi})$ | 0.000419 | 0.000547 | 0.000553 | 0.001071 | 0.006663 |
| Theoretical $Var(\widehat{\pi})$ | 0.000416 | 0.000556 | 0.000567 | 0.001093 | 0.005215 |
| Empirical $Mean(\widehat{\omega})$ | 0.687067 | 0.698150 | 0.700115 | 0.699954 | 0.700461 |
| Empirical $Var(\widehat{\omega})$ | 0.036029 | 0.005501 | 0.000371 | 0.000702 | 0.000654 |
| Theoretical $Var(\widehat{\omega})$ | 0.034358 | 0.005498 | 0.000364 | 0.000717 | 0.000654 |
| Optimal $n_1$ | 742 | 736 | | 647 | |
| Optimal $n_2$ | 258 | 264 | | 353 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.100109 | 0.099997 | 0.099671 | 0.099879 | 0.098861 |
| Empirical $Var(\widehat{\pi})$ | 0.000419 | 0.000515 | 0.000553 | 0.001071 | 0.001460 |
| Theoretical $Var(\widehat{\pi})$ | 0.000416 | 0.000506 | 0.000567 | 0.001093 | 0.001339 |
| Empirical $Mean(\widehat{\omega})$ | 0.687067 | 0.695394 | 0.700115 | 0.699954 | 0.700257 |
| Empirical $Var(\widehat{\omega})$ | 0.036029 | 0.010923 | 0.000371 | 0.000702 | 0.000652 |
| Theoretical $Var(\widehat{\omega})$ | 0.034358 | 0.010664 | 0.000364 | 0.000717 | 0.000654 |
| Optimal $n_1$ | 742 | 735 | | 647 | |
| Optimal $n_2$ | 258 | 265 | | 353 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.100109 | 0.100060 | 0.099671 | 0.099879 | 0.099859 |
| Empirical $Var(\widehat{\pi})$ | 0.000419 | 0.000369 | 0.000553 | 0.001071 | 0.000238 |
| Theoretical $Var(\widehat{\pi})$ | 0.000416 | 0.000375 | 0.000567 | 0.001093 | 0.000233 |
| Empirical $Mean(\widehat{\omega})$ | 0.687067 | 0.687874 | 0.700115 | 0.699954 | 0.699923 |
| Empirical $Var(\widehat{\omega})$ | 0.036029 | 0.096462 | 0.000371 | 0.000702 | 0.000655 |
| Theoretical $Var(\widehat{\omega})$ | 0.034358 | 0.096102 | 0.000364 | 0.000717 | 0.000654 |
| Optimal $n_1$ | 742 | 750 | | 647 | |
| Optimal $n_2$ | 258 | 250 | | 353 | |

* More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

## Table 4. Comparison of Different Two-Stage Model Estimators ($\pi = 0.1$ & $\omega = 0.9$)

| $\pi = 0.1$, $\omega = 0.9$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)[*] | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.100208 | 0.100438 | 0.100451 | 0.100365 | 0.094515 |
| Empirical $Var(\widehat{\pi})$ | 0.000438 | 0.000629 | 0.001153 | 0.001136 | 0.005656 |
| Theoretical $Var(\widehat{\pi})$ | 0.000446 | 0.000635 | 0.001165 | 0.001166 | 0.003606 |
| Empirical $Mean(\widehat{\omega})$ | 0.887436 | 0.898947 | 0.899713 | 0.899507 | 0.900061 |
| Empirical $Var(\widehat{\omega})$ | 0.033923 | 0.002691 | 0.000261 | 0.000526 | 0.000534 |
| Theoretical $Var(\widehat{\omega})$ | 0.032875 | 0.002602 | 0.000259 | 0.000535 | 0.000534 |
| Optimal $n_1$ | 738 | 755 | | 644 | |
| Optimal $n_2$ | 262 | 245 | | 356 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.100208 | 0.100383 | 0.100451 | 0.100365 | 0.488790 |
| Empirical $Var(\widehat{\pi})$ | 0.000438 | 0.000598 | 0.001153 | 0.001136 | 34.2695 |
| Theoretical $Var(\widehat{\pi})$ | 0.000446 | 0.000601 | 0.001165 | 0.001166 | 3.906090 |
| Empirical $Mean(\widehat{\omega})$ | 0.887436 | 0.897482 | 0.899713 | 0.899507 | 0.899625 |
| Empirical $Var(\widehat{\omega})$ | 0.033923 | 0.004855 | 0.000261 | 0.000526 | 0.000520 |
| Theoretical $Var(\widehat{\omega})$ | 0.032875 | 0.004797 | 0.000259 | 0.000535 | 0.000534 |
| Optimal $n_1$ | 738 | 742 | | 644 | |
| Optimal $n_2$ | 262 | 258 | | 356 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.100208 | 0.099773 | 0.100451 | 0.100365 | 0.098965 |
| Empirical $Var(\widehat{\pi})$ | 0.000438 | 0.000543 | 0.001153 | 0.001136 | 0.004735 |
| Theoretical $Var(\widehat{\pi})$ | 0.000446 | 0.000550 | 0.001165 | 0.001166 | 0.004049 |
| Empirical $Mean(\widehat{\omega})$ | 0.887436 | 0.897263 | 0.899713 | 0.899507 | 0.899769 |
| Empirical $Var(\widehat{\omega})$ | 0.033923 | 0.009807 | 0.000261 | 0.000526 | 0.000538 |
| Theoretical $Var(\widehat{\omega})$ | 0.032875 | 0.009897 | 0.000259 | 0.000535 | 0.000534 |
| Optimal $n_1$ | 738 | 735 | | 644 | |
| Optimal $n_2$ | 262 | 265 | | 356 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.100208 | 0.100039 | 0.100451 | 0.100365 | 0.099798 |
| Empirical $Var(\widehat{\pi})$ | 0.000438 | 0.000398 | 0.001153 | 0.001136 | 0.000290 |
| Theoretical $Var(\widehat{\pi})$ | 0.000446 | 0.000396 | 0.001165 | 0.001166 | 0.000292 |
| Empirical $Mean(\widehat{\omega})$ | 0.887436 | 0.887998 | 0.899713 | 0.899507 | 0.899787 |
| Empirical $Var(\widehat{\omega})$ | 0.033923 | 0.098139 | 0.000261 | 0.000526 | 0.000539 |
| Theoretical $Var(\widehat{\omega})$ | 0.032875 | 0.097850 | 0.000259 | 0.000535 | 0.000534 |
| Optimal $n_1$ | 738 | 745 | | 644 | |
| Optimal $n_2$ | 262 | 255 | | 356 | |

[*] More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

Table 5. Comparison of Different Two-Stage Model Estimators ($\pi = 0.3$ & $\omega = 0.1$)

| $\pi = 0.3$, $\omega = 0.1$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)[*] | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300084 | 0.299671 | 0.300164 | 0.299607 | 0.299420 |
| Empirical $Var(\widehat{\pi})$ | 0.000682 | 0.000699 | 0.000243 | 0.001849 | 0.000339 |
| Theoretical $Var(\widehat{\pi})$ | 0.000682 | 0.000696 | 0.000241 | 0.001883 | 0.000335 |
| Empirical $Mean(\widehat{\omega})$ | 7.950570 | 0.083196 | 0.100042 | 0.100060 | 0.100258 |
| Empirical $Var(\widehat{\omega})$ | 676426 | 0.031060 | 0.000193 | 0.000404 | 0.000532 |
| Theoretical $Var(\widehat{\omega})$ | 1.831340 | 0.027656 | 0.000199 | 0.000410 | 0.000534 |
| Optimal $n_1$ | 777 | 775 | | 667 | |
| Optimal $n_2$ | 223 | 225 | | 333 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300084 | 0.300523 | 0.300164 | 0.299607 | 0.299949 |
| Empirical $Var(\widehat{\pi})$ | 0.000682 | 0.000694 | 0.000243 | 0.001849 | 0.000301 |
| Theoretical $Var(\widehat{\pi})$ | 0.000682 | 0.000693 | 0.000241 | 0.001883 | 0.000301 |
| Empirical $Mean(\widehat{\omega})$ | 7.950570 | 0.072380 | 0.100042 | 0.100060 | 0.099802 |
| Empirical $Var(\widehat{\omega})$ | 676426 | 0.050442 | 0.000193 | 0.000404 | 0.000520 |
| Theoretical $Var(\widehat{\omega})$ | 1.831340 | 0.043985 | 0.000199 | 0.000410 | 0.000534 |
| Optimal $n_1$ | 777 | 775 | | 667 | |
| Optimal $n_2$ | 223 | 225 | | 333 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300084 | 0.300057 | 0.300164 | 0.299607 | 0.299879 |
| Empirical $Var(\widehat{\pi})$ | 0.000682 | 0.000690 | 0.000243 | 0.001849 | 0.000261 |
| Theoretical $Var(\widehat{\pi})$ | 0.000682 | 0.000690 | 0.000241 | 0.001883 | 0.000272 |
| Empirical $Mean(\widehat{\omega})$ | 7.950570 | 0.067591 | 0.100042 | 0.100060 | 0.099740 |
| Empirical $Var(\widehat{\omega})$ | 676426 | 0.089037 | 0.000193 | 0.000404 | 0.000533 |
| Theoretical $Var(\widehat{\omega})$ | 1.831340 | 0.079829 | 0.000199 | 0.000410 | 0.000534 |
| Optimal $n_1$ | 777 | 776 | | 667 | |
| Optimal $n_2$ | 223 | 224 | | 333 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300084 | 0.299935 | 0.300164 | 0.299607 | 0.299810 |
| Empirical $Var(\widehat{\pi})$ | 0.000682 | 0.000696 | 0.000243 | 0.001849 | 0.000230 |
| Theoretical $Var(\widehat{\pi})$ | 0.000682 | 0.000684 | 0.000241 | 0.001883 | 0.000226 |
| Empirical $Mean(\widehat{\omega})$ | 7.950570 | 0.004726 | 0.100042 | 0.100060 | 0.099990 |
| Empirical $Var(\widehat{\omega})$ | 676426 | 0.850841 | 0.000193 | 0.000404 | 0.000538 |
| Theoretical $Var(\widehat{\omega})$ | 1.831340 | 0.745353 | 0.000199 | 0.000410 | 0.000534 |
| Optimal $n_1$ | 777 | 777 | | 667 | |
| Optimal $n_2$ | 223 | 223 | | 333 | |

[*] More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

Table 6. Comparison of Different Two-Stage Model Estimators ($\pi = 0.3$ & $\omega = 0.3$)

| $\pi = 0.3$, $\omega = 0.3$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)* | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300065 | 0.299781 | 0.299980 | 0.300410 | 0.299133 |
| Empirical $Var(\widehat{\pi})$ | 0.000688 | 0.000730 | 0.000325 | 0.001874 | 0.000772 |
| Theoretical $Var(\widehat{\pi})$ | 0.000686 | 0.000722 | 0.000330 | 0.001868 | 0.000732 |
| Empirical $Mean(\widehat{\omega})$ | 21.6122 | 0.285857 | 0.300242 | 0.300316 | 0.300022 |
| Empirical $Var(\widehat{\omega})$ | 1036560 | 0.024821 | 0.000327 | 0.000648 | 0.000663 |
| Theoretical $Var(\widehat{\omega})$ | 1.628680 | 0.022387 | 0.000334 | 0.000644 | 0.000654 |
| Optimal $n_1$ | 777 | 771 | | 668 | |
| Optimal $n_2$ | 223 | 229 | | 332 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300065 | 0.299722 | 0.299980 | 0.300410 | 0.299495 |
| Empirical $Var(\widehat{\pi})$ | 0.000688 | 0.000720 | 0.000325 | 0.001874 | 0.000554 |
| Theoretical $Var(\widehat{\pi})$ | 0.000686 | 0.000715 | 0.000330 | 0.001868 | 0.000549 |
| Empirical $Mean(\widehat{\omega})$ | 21.6122 | 0.281282 | 0.300242 | 0.300316 | 0.300136 |
| Empirical $Var(\widehat{\omega})$ | 1036560 | 0.041955 | 0.000327 | 0.000648 | 0.000627 |
| Theoretical $Var(\widehat{\omega})$ | 1.628680 | 0.037382 | 0.000334 | 0.000644 | 0.000654 |
| Optimal $n_1$ | 777 | 772 | | 668 | |
| Optimal $n_2$ | 223 | 228 | | 332 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300065 | 0.299467 | 0.299980 | 0.300410 | 0.299404 |
| Empirical $Var(\widehat{\pi})$ | 0.000688 | 0.000699 | 0.000325 | 0.001874 | 0.000414 |
| Theoretical $Var(\widehat{\pi})$ | 0.000686 | 0.000707 | 0.000330 | 0.001868 | 0.000421 |
| Empirical $Mean(\widehat{\omega})$ | 21.6122 | 0.279218 | 0.300242 | 0.300316 | 0.300740 |
| Empirical $Var(\widehat{\omega})$ | 1036560 | 0.078808 | 0.000327 | 0.000648 | 0.000648 |
| Theoretical $Var(\widehat{\omega})$ | 1.628680 | 0.070802 | 0.000334 | 0.000644 | 0.000654 |
| Optimal $n_1$ | 777 | 773 | | 668 | |
| Optimal $n_2$ | 223 | 227 | | 332 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300065 | 0.299814 | 0.299980 | 0.300410 | 0.299984 |
| Empirical $Var(\widehat{\pi})$ | 0.000688 | 0.000694 | 0.000325 | 0.001874 | 0.000258 |
| Theoretical $Var(\widehat{\pi})$ | 0.000686 | 0.000690 | 0.000330 | 0.001868 | 0.000260 |
| Empirical $Mean(\widehat{\omega})$ | 21.6122 | 0.207572 | 0.300242 | 0.300316 | 0.299946 |
| Empirical $Var(\widehat{\omega})$ | 1036560 | 0.805448 | 0.000327 | 0.000648 | 0.000652 |
| Theoretical $Var(\widehat{\omega})$ | 1.628680 | 0.718458 | 0.000334 | 0.000644 | 0.000654 |
| Optimal $n_1$ | 777 | 776 | | 668 | |
| Optimal $n_2$ | 223 | 224 | | 332 | |

* More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

Table 7. Comparison of Different Two-Stage Model Estimators ($\pi = 0.3$ & $\omega = 0.7$)

| $\pi = 0.3$, $\omega = 0.7$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)* | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.299816 | 0.300042 | 0.300132 | 0.300144 | 0.450579 |
| Empirical $Var(\widehat{\pi})$ | 0.000691 | 0.000768 | 0.000756 | 0.001854 | 10.724 |
| Theoretical $Var(\widehat{\pi})$ | 0.000694 | 0.000757 | 0.000769 | 0.001838 | 0.624960 |
| Empirical $Mean(\widehat{\omega})$ | 0.773357 | 0.692130 | 0.699766 | 0.700176 | 0.699565 |
| Empirical $Var(\widehat{\omega})$ | 259.314 | 0.015686 | 0.000368 | 0.000721 | 0.000660 |
| Theoretical $Var(\widehat{\omega})$ | 1.2671 | 0.014089 | 0.000364 | 0.000701 | 0.000654 |
| Optimal $n_1$ | 775 | 770 | | 669 | |
| Optimal $n_2$ | 225 | 230 | | 331 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.299816 | 0.299547 | 0.300132 | 0.300144 | 0.297795 |
| Empirical $Var(\widehat{\pi})$ | 0.000691 | 0.000743 | 0.000756 | 0.001854 | 0.005868 |
| Theoretical $Var(\widehat{\pi})$ | 0.000694 | 0.000748 | 0.000769 | 0.001838 | 0.005323 |
| Empirical $Mean(\widehat{\omega})$ | 0.773357 | 0.686035 | 0.699766 | 0.700176 | 0.700143 |
| Empirical $Var(\widehat{\omega})$ | 259.314 | 0.028331 | 0.000368 | 0.000721 | 0.000660 |
| Theoretical $Var(\widehat{\omega})$ | 1.2671 | 0.025912 | 0.000364 | 0.000701 | 0.000654 |
| Optimal $n_1$ | 775 | 770 | | 669 | |
| Optimal $n_2$ | 225 | 230 | | 331 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.299816 | 0.300178 | 0.300132 | 0.300144 | 0.299264 |
| Empirical $Var(\widehat{\pi})$ | 0.000691 | 0.000709 | 0.000756 | 0.001854 | 0.001471 |
| Theoretical $Var(\widehat{\pi})$ | 0.000694 | 0.000735 | 0.000769 | 0.001838 | 0.001439 |
| Empirical $Mean(\widehat{\omega})$ | 0.773357 | 0.673376 | 0.699766 | 0.700176 | 0.700072 |
| Empirical $Var(\widehat{\omega})$ | 259.314 | 0.060151 | 0.000368 | 0.000721 | 0.000650 |
| Theoretical $Var(\widehat{\omega})$ | 1.2671 | 0.054235 | 0.000364 | 0.000701 | 0.000654 |
| Optimal $n_1$ | 775 | 770 | | 669 | |
| Optimal $n_2$ | 225 | 230 | | 331 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.299816 | 0.300006 | 0.300132 | 0.300144 | 0.299604 |
| Empirical $Var(\widehat{\pi})$ | 0.000691 | 0.000712 | 0.000756 | 0.001854 | 0.000352 |
| Theoretical $Var(\widehat{\pi})$ | 0.000694 | 0.000701 | 0.000769 | 0.001838 | 0.000348 |
| Empirical $Mean(\widehat{\omega})$ | 0.773357 | 0.606874 | 0.699766 | 0.700176 | 0.700101 |
| Empirical $Var(\widehat{\omega})$ | 259.314 | 0.762194 | 0.000368 | 0.000721 | 0.000652 |
| Theoretical $Var(\widehat{\omega})$ | 1.2671 | 0.664273 | 0.000364 | 0.000701 | 0.000654 |
| Optimal $n_1$ | 775 | 774 | | 669 | |
| Optimal $n_2$ | 225 | 226 | | 331 | |

* More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

## Table 8. Comparison of Different Two-Stage Model Estimators ($\pi = 0.3$ & $\omega = 0.9$)

| $\pi = 0.3$, $\omega = 0.9$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)* | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300066 | 0.300121 | 0.300199 | 0.300339 | 0.297836 |
| Empirical $Var(\widehat{\pi})$ | 0.000680 | 0.000779 | 0.001449 | 0.001813 | 0.004313 |
| Theoretical $Var(\widehat{\pi})$ | 0.000698 | 0.000767 | 0.001437 | 0.001822 | 0.003675 |
| Empirical $Mean(\widehat{\omega})$ | 0.869297 | 0.896474 | 0.89995 | 0.899966 | 0.900186 |
| Empirical $Var(\widehat{\omega})$ | 183.953 | 0.012476 | 0.000262 | 0.000517 | 0.000532 |
| Theoretical $Var(\widehat{\omega})$ | 1.11997 | 0.011666 | 0.000259 | 0.000524 | 0.000534 |
| Optimal $n_1$ | 775 | 773 | | 670 | |
| Optimal $n_2$ | 225 | 227 | | 330 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300066 | 0.299509 | 0.300199 | 0.300339 | 0.434294 |
| Empirical $Var(\widehat{\pi})$ | 0.000680 | 0.000746 | 0.001449 | 0.001813 | 29.2759 |
| Theoretical $Var(\widehat{\pi})$ | 0.000698 | 0.000759 | 0.001437 | 0.001822 | 3.90621 |
| Empirical $Mean(\widehat{\omega})$ | 0.869297 | 0.889245 | 0.899950 | 0.899966 | 0.899741 |
| Empirical $Var(\widehat{\omega})$ | 183.953 | 0.023423 | 0.000262 | 0.000517 | 0.000533 |
| Theoretical $Var(\widehat{\omega})$ | 1.11997 | 0.021525 | 0.000259 | 0.000524 | 0.000534 |
| Optimal $n_1$ | 775 | 770 | | 670 | |
| Optimal $n_2$ | 225 | 230 | | 330 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300066 | 0.300539 | 0.300199 | 0.300339 | 0.298900 |
| Empirical $Var(\widehat{\pi})$ | 0.000680 | 0.000740 | 0.001449 | 0.001813 | 0.004425 |
| Theoretical $Var(\widehat{\pi})$ | 0.000698 | 0.000746 | 0.001437 | 0.001822 | 0.004162 |
| Empirical $Mean(\widehat{\omega})$ | 0.869297 | 0.875158 | 0.899950 | 0.899966 | 0.900026 |
| Empirical $Var(\widehat{\omega})$ | 183.953 | 0.052949 | 0.000262 | 0.000517 | 0.000536 |
| Theoretical $Var(\widehat{\omega})$ | 1.11997 | 0.047165 | 0.000259 | 0.000524 | 0.000534 |
| Optimal $n_1$ | 775 | 770 | | 670 | |
| Optimal $n_2$ | 225 | 230 | | 330 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.300066 | 0.300362 | 0.300199 | 0.300339 | 0.299860 |
| Empirical $Var(\widehat{\pi})$ | 0.000680 | 0.000723 | 0.001449 | 0.001813 | 0.000407 |
| Theoretical $Var(\widehat{\pi})$ | 0.000698 | 0.000707 | 0.001437 | 0.001822 | 0.000408 |
| Empirical $Mean(\widehat{\omega})$ | 0.869297 | 0.799196 | 0.899950 | 0.899966 | 0.899706 |
| Empirical $Var(\widehat{\omega})$ | 183.953 | 0.737917 | 0.000262 | 0.000517 | 0.000548 |
| Theoretical $Var(\widehat{\omega})$ | 1.11997 | 0.637221 | 0.000259 | 0.000524 | 0.000534 |
| Optimal $n_1$ | 775 | 773 | | 670 | |
| Optimal $n_2$ | 225 | 227 | | 330 | |

* More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

## Table 9. Comparison of Different Two-Stage Model Estimators ($\pi = 0.6$ & $\omega = 0.1$)

| $\pi = 0.6$, $\omega = 0.1$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)* | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599931 | 0.600114 | 0.600047 | 0.600064 | 0.599969 |
| Empirical $Var(\widehat{\pi})$ | 0.000782 | 0.000796 | 0.000291 | 0.002192 | 0.000346 |
| Theoretical $Var(\widehat{\pi})$ | 0.000782 | 0.000781 | 0.000296 | 0.002183 | 0.000337 |
| Empirical $Mean(\widehat{\omega})$ | 0.073193 | -0.011230 | 0.099866 | 0.100070 | 0.099984 |
| Empirical $Var(\widehat{\omega})$ | 0.090768 | 0.835511 | 0.000203 | 0.000409 | 0.000524 |
| Theoretical $Var(\widehat{\omega})$ | 0.084199 | 0.123181 | 0.000199 | 0.000410 | 0.000534 |
| Optimal $n_1$ | 777 | 777 | | 666 | |
| Optimal $n_2$ | 223 | 223 | | 334 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599931 | 0.600557 | 0.600047 | 0.600064 | 0.599706 |
| Empirical $Var(\widehat{\pi})$ | 0.000782 | 0.000790 | 0.000291 | 0.002192 | 0.000321 |
| Theoretical $Var(\widehat{\pi})$ | 0.000782 | 0.000781 | 0.000296 | 0.002183 | 0.000313 |
| Empirical $Mean(\widehat{\omega})$ | 0.073193 | -0.036102 | 0.099866 | 0.100070 | 0.099784 |
| Empirical $Var(\widehat{\omega})$ | 0.090768 | 1.90469 | 0.000203 | 0.000409 | 0.000542 |
| Theoretical $Var(\widehat{\omega})$ | 0.084199 | 0.197117 | 0.000199 | 0.000410 | 0.000534 |
| Optimal $n_1$ | 777 | 777 | | 666 | |
| Optimal $n_2$ | 223 | 223 | | 334 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599931 | 0.600140 | 0.600047 | 0.600064 | 0.600099 |
| Empirical $Var(\widehat{\pi})$ | 0.000782 | 0.000769 | 0.000291 | 0.002192 | 0.000292 |
| Theoretical $Var(\widehat{\pi})$ | 0.000782 | 0.000780 | 0.000296 | 0.002183 | 0.000291 |
| Empirical $Mean(\widehat{\omega})$ | 0.073193 | -0.093077 | 0.099866 | 0.100070 | 0.099774 |
| Empirical $Var(\widehat{\omega})$ | 0.090768 | 1.12045 | 0.000203 | 0.000409 | 0.000537 |
| Theoretical $Var(\widehat{\omega})$ | 0.084199 | 0.358819 | 0.000199 | 0.000410 | 0.000534 |
| Optimal $n_1$ | 777 | 777 | | 666 | |
| Optimal $n_2$ | 223 | 223 | | 334 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599931 | 0.599834 | 0.600047 | 0.600064 | 0.599954 |
| Empirical $Var(\widehat{\pi})$ | 0.000782 | 0.000777 | 0.000291 | 0.002192 | 0.000258 |
| Theoretical $Var(\widehat{\pi})$ | 0.000782 | 0.000778 | 0.000296 | 0.002183 | 0.000255 |
| Empirical $Mean(\widehat{\omega})$ | 0.073193 | -0.523518 | 0.099866 | 0.100070 | 0.099937 |
| Empirical $Var(\widehat{\omega})$ | 0.090768 | 9.93607 | 0.000203 | 0.000409 | 0.000541 |
| Theoretical $Var(\widehat{\omega})$ | 0.084199 | 3.39486 | 0.000199 | 0.000410 | 0.000534 |
| Optimal $n_1$ | 777 | 778 | | 666 | |
| Optimal $n_2$ | 223 | 222 | | 334 | |

* More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

## Table 10. Comparison of Different Two-Stage Model Estimators ($\pi = 0.6$ & $\omega = 0.3$)

| $\pi = 0.6$, $\omega = 0.3$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)[*] | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599892 | 0.599860 | 0.600313 | 0.600606 | 0.599877 |
| Empirical $Var(\widehat{\pi})$ | 0.000793 | 0.000779 | 0.000389 | 0.002174 | 0.000750 |
| Theoretical $Var(\widehat{\pi})$ | 0.000790 | 0.000788 | 0.000432 | 0.002218 | 0.000740 |
| Empirical $Mean(\widehat{\omega})$ | 0.273981 | 0.184471 | 0.300046 | 0.300003 | 0.300398 |
| Empirical $Var(\widehat{\omega})$ | 0.083689 | 3.46138 | 0.000335 | 0.000658 | 0.000648 |
| Theoretical $Var(\widehat{\omega})$ | 0.075041 | 0.095768 | 0.000334 | 0.000646 | 0.000654 |
| Optimal $n_1$ | 776 | 776 | | 665 | |
| Optimal $n_2$ | 224 | 224 | | 335 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599892 | 0.599457 | 0.600313 | 0.600606 | 0.599990 |
| Empirical $Var(\widehat{\pi})$ | 0.000793 | 0.000810 | 0.000389 | 0.002174 | 0.000568 |
| Theoretical $Var(\widehat{\pi})$ | 0.000790 | 0.000786 | 0.000432 | 0.002218 | 0.000562 |
| Empirical $Mean(\widehat{\omega})$ | 0.273981 | 0.143794 | 0.300046 | 0.300003 | 0.299786 |
| Empirical $Var(\widehat{\omega})$ | 0.083689 | 1.27267 | 0.000335 | 0.000658 | 0.000644 |
| Theoretical $Var(\widehat{\omega})$ | 0.075041 | 0.161466 | 0.000334 | 0.000646 | 0.000654 |
| Optimal $n_1$ | 776 | 776 | | 665 | |
| Optimal $n_2$ | 224 | 224 | | 335 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599892 | 0.599191 | 0.600313 | 0.600606 | 0.600314 |
| Empirical $Var(\widehat{\pi})$ | 0.000793 | 0.000797 | 0.000389 | 0.002174 | 0.000451 |
| Theoretical $Var(\widehat{\pi})$ | 0.000790 | 0.000784 | 0.000432 | 0.002218 | 0.000440 |
| Empirical $Mean(\widehat{\omega})$ | 0.273981 | 0.037344 | 0.300046 | 0.300003 | 0.300156 |
| Empirical $Var(\widehat{\omega})$ | 0.083689 | 45.7886 | 0.000335 | 0.000658 | 0.000652 |
| Theoretical $Var(\widehat{\omega})$ | 0.075041 | 0.310463 | 0.000334 | 0.000646 | 0.000654 |
| Optimal $n_1$ | 776 | 777 | | 665 | |
| Optimal $n_2$ | 224 | 223 | | 335 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599892 | 0.600618 | 0.600313 | 0.600606 | 0.599917 |
| Empirical $Var(\widehat{\pi})$ | 0.000793 | 0.000775 | 0.000389 | 0.002174 | 0.000290 |
| Theoretical $Var(\widehat{\pi})$ | 0.000790 | 0.000780 | 0.000432 | 0.002218 | 0.000288 |
| Empirical $Mean(\widehat{\omega})$ | 0.273981 | -0.170965 | 0.300046 | 0.300003 | 0.300031 |
| Empirical $Var(\widehat{\omega})$ | 0.083689 | 35.6447 | 0.000335 | 0.000658 | 0.000665 |
| Theoretical $Var(\widehat{\omega})$ | 0.075041 | 3.22937 | 0.000334 | 0.000646 | 0.000654 |
| Optimal $n_1$ | 776 | 777 | | 665 | |
| Optimal $n_2$ | 224 | 223 | | 335 | |

[*] More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

Table 11. Comparison of Different Two-Stage Model Estimators ($\pi = 0.6$ & $\omega = 0.7$)

| $\pi = 0.6$, $\omega = 0.7$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)* | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599905 | 0.600216 | 0.599866 | 0.599672 | 0.512808 |
| Empirical $Var(\widehat{\pi})$ | 0.000782 | 0.000809 | 0.000922 | 0.002242 | 11.3447 |
| Theoretical $Var(\widehat{\pi})$ | 0.000799 | 0.000797 | 0.001026 | 0.002239 | 0.624990 |
| Empirical $Mean(\widehat{\omega})$ | 0.676777 | 0.651850 | 0.699945 | 0.699750 | 0.699687 |
| Empirical $Var(\widehat{\omega})$ | 0.062474 | 0.194263 | 0.000363 | 0.000706 | 0.000662 |
| Theoretical $Var(\widehat{\omega})$ | 0.057994 | 0.057869 | 0.000364 | 0.000703 | 0.000654 |
| Optimal $n_1$ | 776 | 776 | | 667 | |
| Optimal $n_2$ | 224 | 224 | | 333 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599905 | 0.600202 | 0.599866 | 0.599672 | 0.602554 |
| Empirical $Var(\widehat{\pi})$ | 0.000782 | 0.000784 | 0.000922 | 0.002242 | 0.005694 |
| Theoretical $Var(\widehat{\pi})$ | 0.000799 | 0.000794 | 0.001026 | 0.002239 | 0.005349 |
| Empirical $Mean(\widehat{\omega})$ | 0.676777 | 0.626695 | 0.699945 | 0.699750 | 0.699946 |
| Empirical $Var(\widehat{\omega})$ | 0.062474 | 0.245408 | 0.000363 | 0.000706 | 0.000657 |
| Theoretical $Var(\widehat{\omega})$ | 0.057994 | 0.107107 | 0.000364 | 0.000703 | 0.000654 |
| Optimal $n_1$ | 776 | 776 | | 667 | |
| Optimal $n_2$ | 224 | 224 | | 333 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599905 | 0.600428 | 0.599866 | 0.599672 | 0.599984 |
| Empirical $Var(\widehat{\pi})$ | 0.000782 | 0.000780 | 0.000922 | 0.002242 | 0.001490 |
| Theoretical $Var(\widehat{\pi})$ | 0.000799 | 0.000791 | 0.001026 | 0.002239 | 0.001464 |
| Empirical $Mean(\widehat{\omega})$ | 0.676777 | 0.548970 | 0.699945 | 0.699750 | 0.699953 |
| Empirical $Var(\widehat{\omega})$ | 0.062474 | 8.959240 | 0.000363 | 0.000706 | 0.000656 |
| Theoretical $Var(\widehat{\omega})$ | 0.057994 | 0.227916 | 0.000364 | 0.000703 | 0.000654 |
| Optimal $n_1$ | 776 | 776 | | 667 | |
| Optimal $n_2$ | 224 | 224 | | 333 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599905 | 0.600202 | 0.599866 | 0.599672 | 0.600105 |
| Empirical $Var(\widehat{\pi})$ | 0.000782 | 0.000777 | 0.000922 | 0.002242 | 0.000380 |
| Theoretical $Var(\widehat{\pi})$ | 0.000799 | 0.000783 | 0.001026 | 0.002239 | 0.000377 |
| Empirical $Mean(\widehat{\omega})$ | 0.676777 | 0.189206 | 0.699945 | 0.699750 | 0.699544 |
| Empirical $Var(\widehat{\omega})$ | 0.062474 | 14.7166 | 0.000363 | 0.000706 | 0.000664 |
| Theoretical $Var(\widehat{\omega})$ | 0.057994 | 2.93443 | 0.000364 | 0.000703 | 0.000654 |
| Optimal $n_1$ | 776 | 777 | | 667 | |
| Optimal $n_2$ | 224 | 223 | | 333 | |

* More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

Table 12. Comparison of Different Two-Stage Model Estimators ($\pi = 0.6$ & $\omega = 0.9$)

| $\pi = 0.6$, $\omega = 0.9$ | Gupta, Tuck, Spears Gill, and Crowe (2013) | Sihm and Gupta (2015) | Sihm, Chhabra, and Gupta (2016) | Revised Sihm, Chhabra, and Gupta (2016)[*] | Proposed § 4.4 |
|---|---|---|---|---|---|
| | | $T = 0.0$ | | | $T = 0.0$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599913 | 0.599612 | 0.599729 | 0.600570 | 0.601235 |
| Empirical $Var(\widehat{\pi})$ | 0.000797 | 0.000796 | 0.001743 | 0.002205 | 0.004068 |
| Theoretical $Var(\widehat{\pi})$ | 0.000800 | 0.000799 | 0.001834 | 0.002226 | 0.003692 |
| Empirical $Mean(\widehat{\omega})$ | 0.882876 | 0.873725 | 0.899855 | 0.899930 | 0.899635 |
| Empirical $Var(\widehat{\omega})$ | 0.054863 | 0.110643 | 0.000261 | 0.000528 | 0.000527 |
| Theoretical $Var(\widehat{\omega})$ | 0.050602 | 0.047935 | 0.000259 | 0.000524 | 0.000534 |
| Optimal $n_1$ | 778 | 777 | | 670 | |
| Optimal $n_2$ | 222 | 223 | | 330 | |
| | | $T = 0.2$ | | | $T = 0.2$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599913 | 0.599793 | 0.599729 | 0.600570 | 0.559286 |
| Empirical $Var(\widehat{\pi})$ | 0.000797 | 0.000790 | 0.001743 | 0.002205 | 36.194100 |
| Theoretical $Var(\widehat{\pi})$ | 0.000800 | 0.000797 | 0.001834 | 0.002226 | 3.906240 |
| Empirical $Mean(\widehat{\omega})$ | 0.882876 | 0.840707 | 0.899855 | 0.899930 | 0.900271 |
| Empirical $Var(\widehat{\omega})$ | 0.054863 | 0.209650 | 0.000261 | 0.000528 | 0.000535 |
| Theoretical $Var(\widehat{\omega})$ | 0.050602 | 0.088407 | 0.000259 | 0.000524 | 0.000534 |
| Optimal $n_1$ | 778 | 776 | | 670 | |
| Optimal $n_2$ | 222 | 224 | | 330 | |
| | | $T = 0.4$ | | | $T = 0.4$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599913 | 0.600034 | 0.599729 | 0.600570 | 0.601268 |
| Empirical $Var(\widehat{\pi})$ | 0.000797 | 0.000799 | 0.001743 | 0.002205 | 0.004304 |
| Theoretical $Var(\widehat{\pi})$ | 0.000800 | 0.000794 | 0.001834 | 0.002226 | 0.004190 |
| Empirical $Mean(\widehat{\omega})$ | 0.882876 | 0.784888 | 0.899855 | 0.899930 | 0.899882 |
| Empirical $Var(\widehat{\omega})$ | 0.054863 | 0.927739 | 0.000261 | 0.000528 | 0.000528 |
| Theoretical $Var(\widehat{\omega})$ | 0.050602 | 0.195299 | 0.000259 | 0.000524 | 0.000534 |
| Optimal $n_1$ | 778 | 776 | | 670 | |
| Optimal $n_2$ | 222 | 224 | | 330 | |
| | | $T = 0.8$ | | | $T = 0.8$ |
| Empirical $Mean(\widehat{\pi})$ | 0.599913 | 0.599697 | 0.599729 | 0.600570 | 0.599836 |
| Empirical $Var(\widehat{\pi})$ | 0.000797 | 0.000787 | 0.001743 | 0.002205 | 0.000439 |
| Theoretical $Var(\widehat{\pi})$ | 0.000800 | 0.000784 | 0.001834 | 0.002226 | 0.000437 |
| Empirical $Mean(\widehat{\omega})$ | 0.882876 | 0.369508 | 0.899855 | 0.899930 | 0.900260 |
| Empirical $Var(\widehat{\omega})$ | 0.054863 | 9.55863 | 0.000261 | 0.000528 | 0.000524 |
| Theoretical $Var(\widehat{\omega})$ | 0.050602 | 2.79416 | 0.000259 | 0.000524 | 0.000534 |
| Optimal $n_1$ | 778 | 777 | | 670 | |
| Optimal $n_2$ | 222 | 223 | | 330 | |

[*] More realistic model than Sihm, Chhabra, and Gupta (2016) with unknown innocuous characteristic $\pi_a$ and $\pi_b$.
$p_{a_1} = 0.8$, $p_{a_2} = 0.2$, $p_{b_1} = 0.7$, $p_{b_2} = 0.4$, $p_a = 0.8$, $p_b = 0.3$, $n = 1,000$, $\pi_a = 0.35$, and $\pi_b = 0.25$

### 5.3 Simulation to Verify Suitable $T$ versus $p_a$

Solving the quartic inequality of $Var\left(\widehat{\pi}_p\right) < Var(\widehat{\pi}_r)$ algebraically with respect to $T$ would be complex and time consuming. But if our purpose is simply to make sure we get improved efficiency by adopting a new model, then verifying it and determining suitable intervals of our parameters wouldn't be hard.

In this simulation study, we show that by choosing suitable values of $T$ and $p_a$, we are able to achieve improved efficiency over the revised model of Sihm, Chhabra, and Gupta (2016) in Section 4.3. In Figures 1 - 12, the red colored area is where $Var\left(\widehat{\pi}_p\right) < Var(\widehat{\pi}_r)$ holds true while the black is the opposite. As we set $p_a = 0.8$ in our simulation study in Section 5.1, the horizontal line of $p_a = 0.8$ is added to every figure to indicate the interval of $T$ where $Var\left(\widehat{\pi}_p\right) < Var(\widehat{\pi}_r)$ holds true.

In Figures 1 and 2, we can tell that any value of $T$ will make $Var\left(\widehat{\pi}_p\right)$ smaller than $Var(\widehat{\pi}_r)$ when $p_a = 0.8$. And this observation nicely matches with the simulation study carried out in Tables 1 and 2. In Figure 3, $T$ greater than 0.5 will lead us to smaller value of $Var\left(\widehat{\pi}_p\right)$ than $Var(\widehat{\pi}_r)$ when $p_a = 0.8$. This agrees with the simulation study done in Tables 3. From Figure 4, we can tell that $T$ greater than 0.6 will lead to a smaller value of $Var\left(\widehat{\pi}_p\right)$ than that of $Var(\widehat{\pi}_r)$ when $p_a = 0.8$. This is in accordance with the simulation study in Table 4.

In Figures 5 and 6, we can tell that any value of $T$ will make $Var\left(\widehat{\pi}_p\right)$ smaller than $Var(\widehat{\pi}_r)$ when $p_a = 0.8$. And this observation nicely matches with the simulation study carried out in Tables 5 and 6. In Figure 7, $T$ greater than 0.4 will make $Var\left(\widehat{\pi}_p\right)$ smaller than $Var(\widehat{\pi}_r)$ when $p_a = 0.8$. This matches with the simulation study done in Table 7. From Figure 8, we can tell that $T$ greater than 0.6 will lead to a smaller value of $Var\left(\widehat{\pi}_p\right)$ than that of $Var(\widehat{\pi}_r)$ when $p_a = 0.8$. This is in accordance with the simulation study in Table 8.

In Figures 9 and 10, we can tell that any value of $T$ will make $Var\left(\widehat{\pi}_p\right)$ smaller than $Var(\widehat{\pi}_r)$ when $p_a = 0.8$. And this observation nicely matches with the simulation study carried out in Tables 9 and 10. In Figure 11, $T$ greater than 0.4 will make $Var\left(\widehat{\pi}_p\right)$ smaller than $Var(\widehat{\pi}_r)$ when $p_a = 0.8$. This matches with the simulation study done in Table 11. In Figure 12, $T$ greater than 0.5 will surely make $Var\left(\widehat{\pi}_p\right)$ smaller than $Var(\widehat{\pi}_r)$ when $p_a = 0.8$. This also matches with the simulation study done in Table 12.

In sum, by setting the value of $T$ greater than 0.6, we will get smaller $Var\left(\widehat{\pi}_p\right)$ value than $Var(\widehat{\pi}_r)$, provided that the true level of sensitivity is no greater than 0.9 and the prevalence of the sensitive characteristic in the population is between 0.1 and 0.6.

In this section, we showed that we can select a suitable value of $T$ to achieve better efficiency for our proposed model than the revised model of Sihm, Chhabra, and Gupta (2016) in Section 4.3. By choosing the value of $T$ in the red colored area from Figures 1 - 12, one can always make $Var\left(\widehat{\pi}_p\right)$ smaller than $Var(\widehat{\pi}_r)$. Notice that when different values are assigned to the parameters, one needs to carry out this part of simulation again to get new Figures and choose suitable intervals of $T$ subsequently.

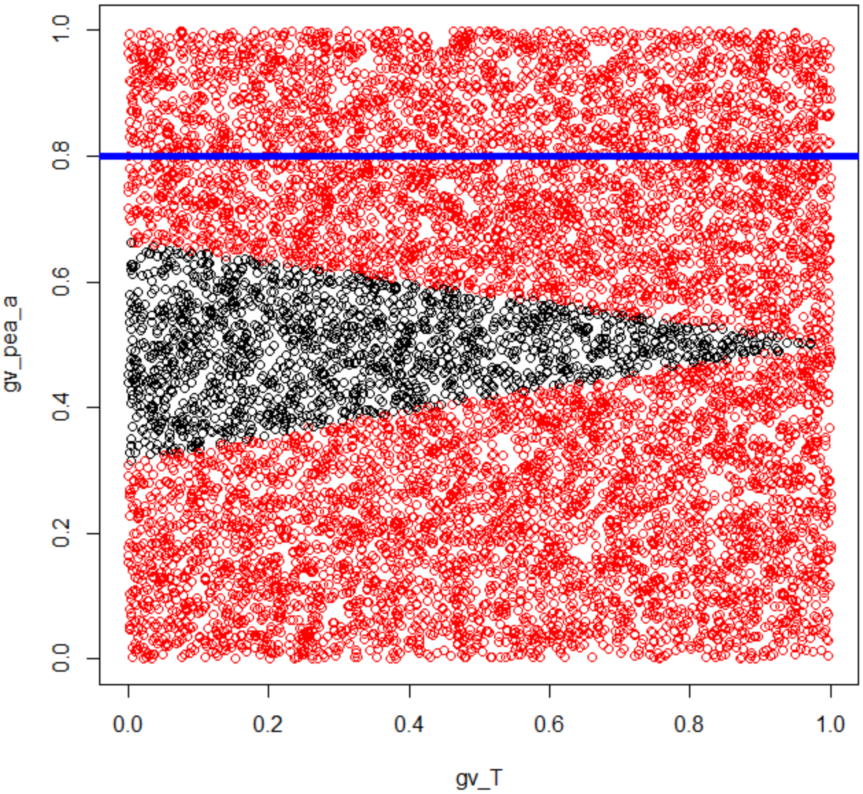

Figure 1. Suitable $T$ Interval for Table 1 ($\pi = 0.1$, $\omega = 0.1$)

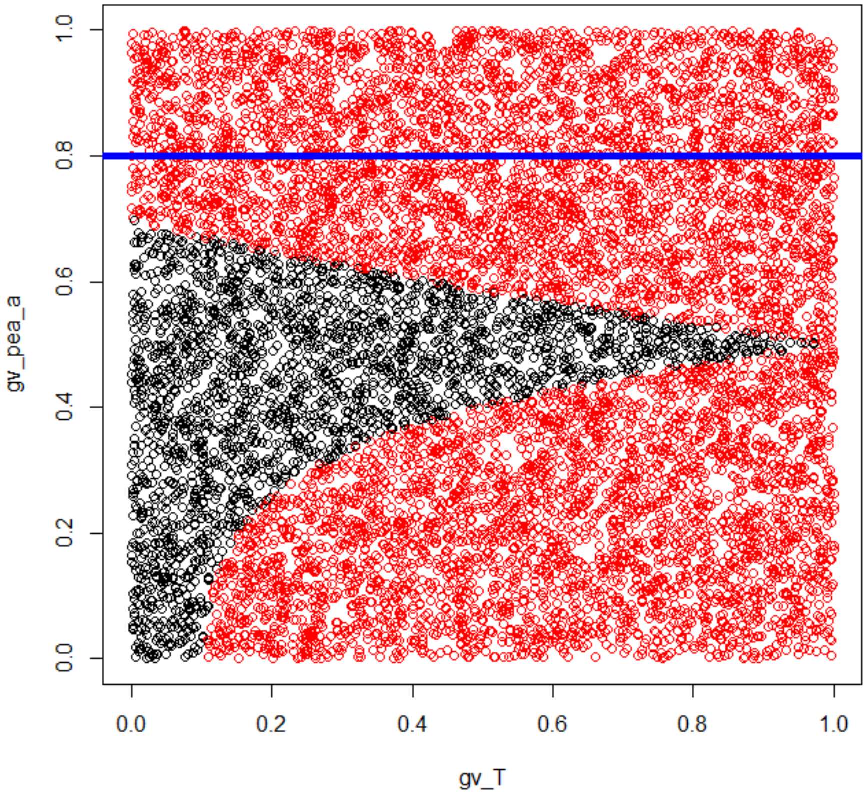

Figure 2. Suitable $T$ Interval for Table 2 ($\pi = 0.1$, $\omega = 0.3$)

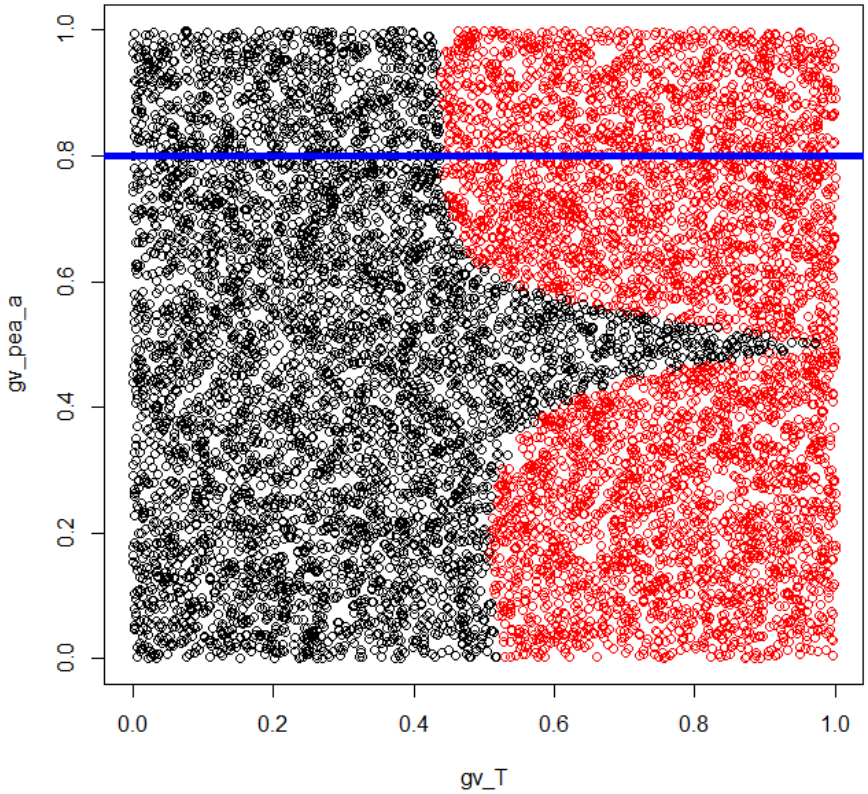

Figure 3. Suitable $T$ Interval for Table 3 ($\pi = 0.1$, $\omega = 0.7$)

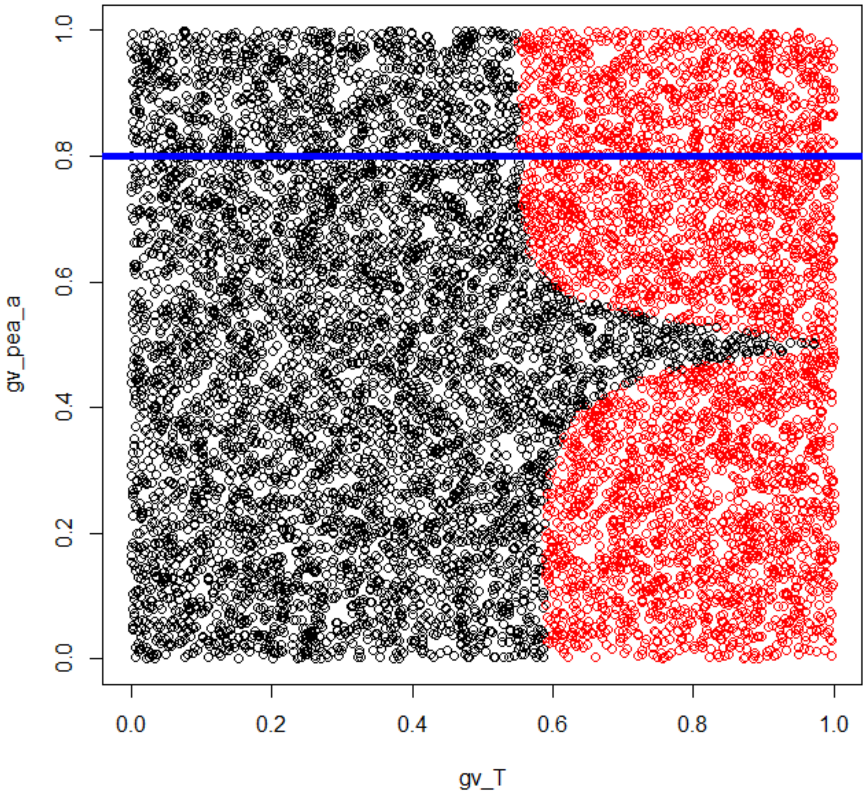

Figure 4. Suitable $T$ Interval for Table 4 ($\pi = 0.1$, $\omega = 0.9$)

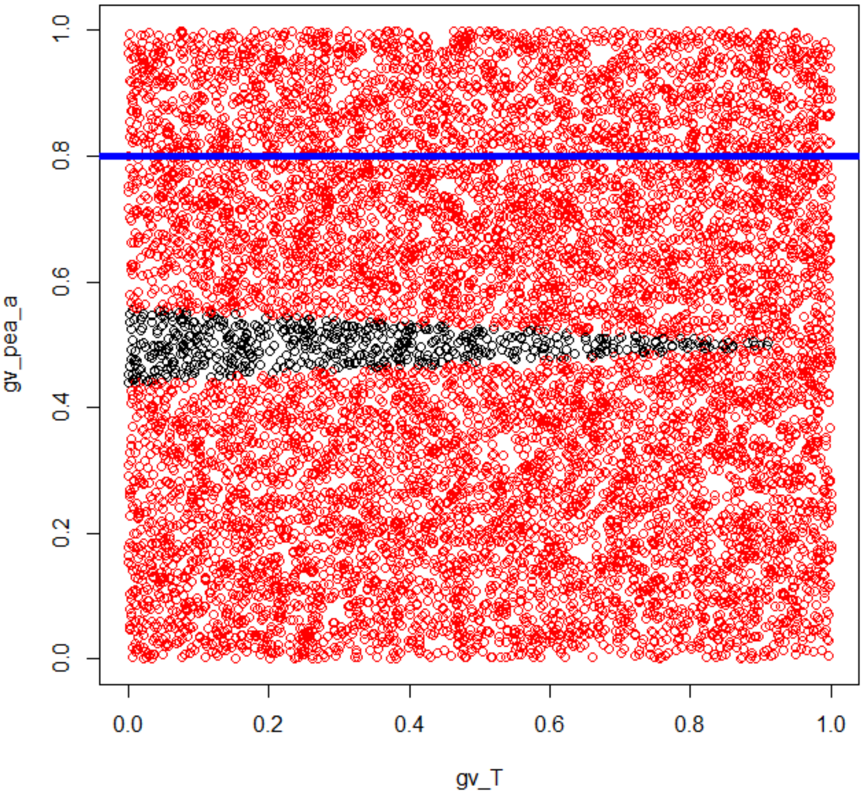

Figure 5. Suitable $T$ Interval for Table 5 ($\pi = 0.3$, $\omega = 0.1$)

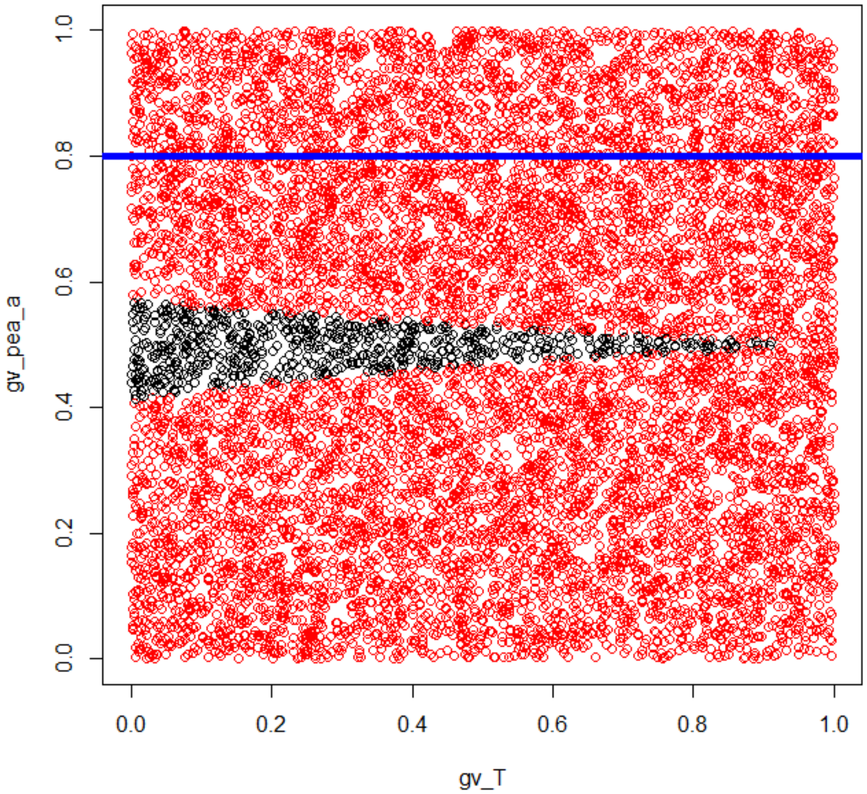

Figure 6. Suitable $T$ Interval for Table 6 ($\pi = 0.3$, $\omega = 0.3$)

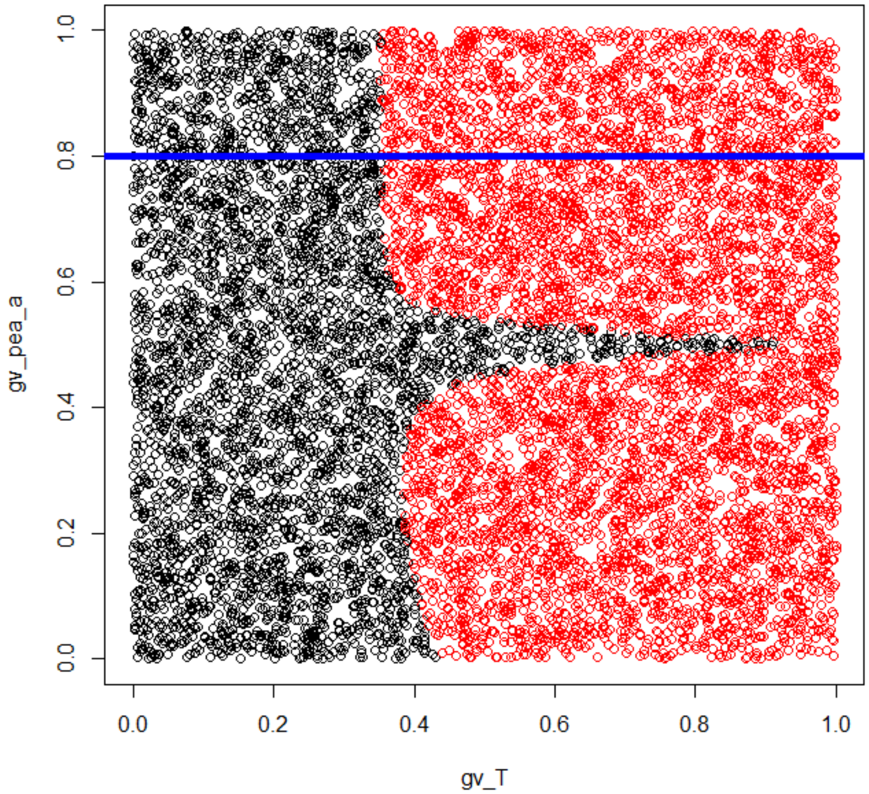

Figure 7. Suitable $T$ Interval for Table 7 ($\pi = 0.3$, $\omega = 0.7$)

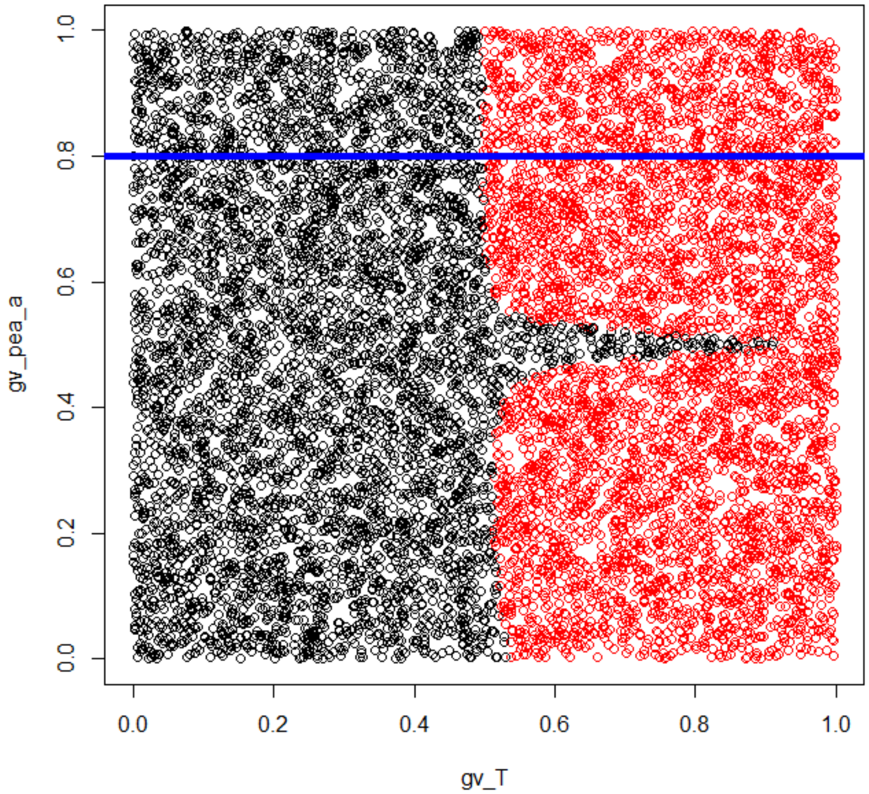

Figure 8. Suitable $T$ Interval for Table 8 ($\pi = 0.3$, $\omega = 0.9$)

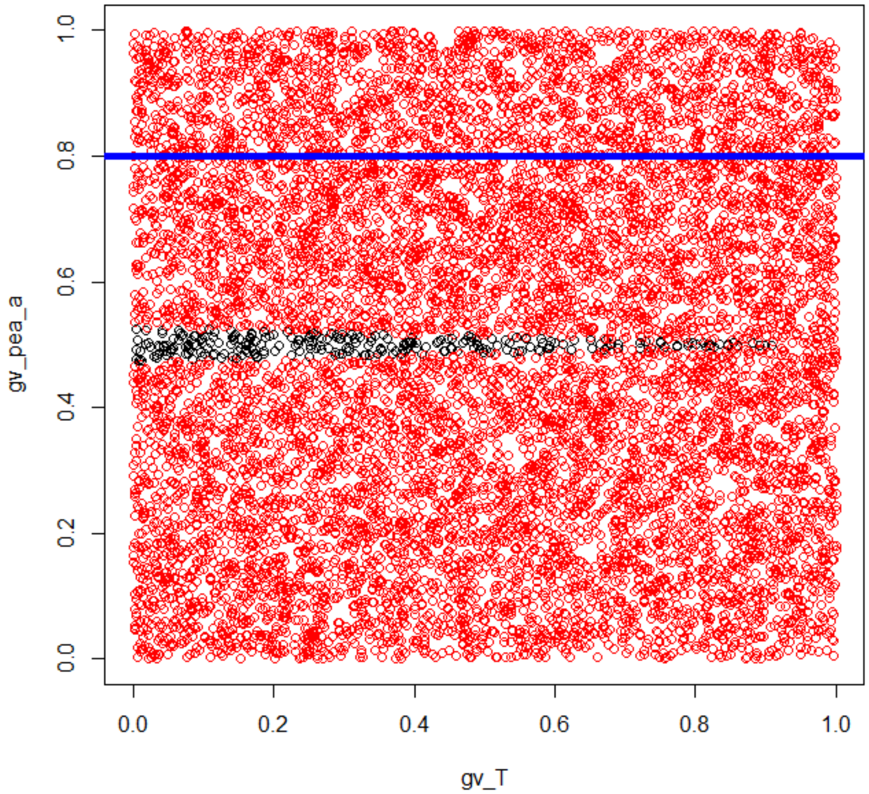

Figure 9. Suitable $T$ Interval for Table 9 ($\pi = 0.6$, $\omega = 0.1$)

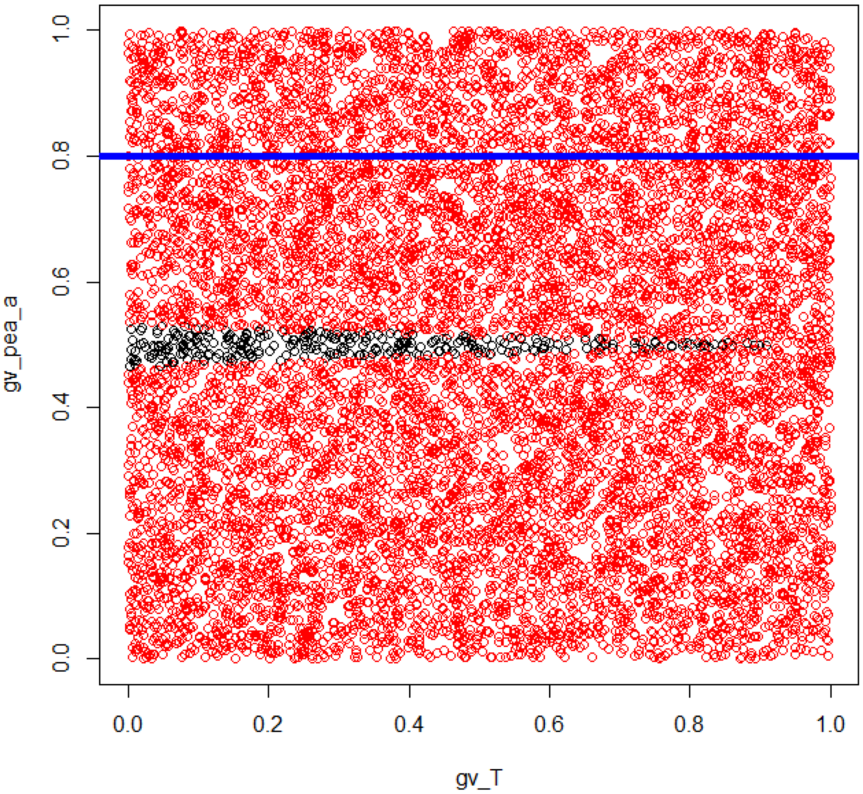

Figure 10. Suitable $T$ Interval for Table 10 ($\pi = 0.6$, $\omega = 0.3$)

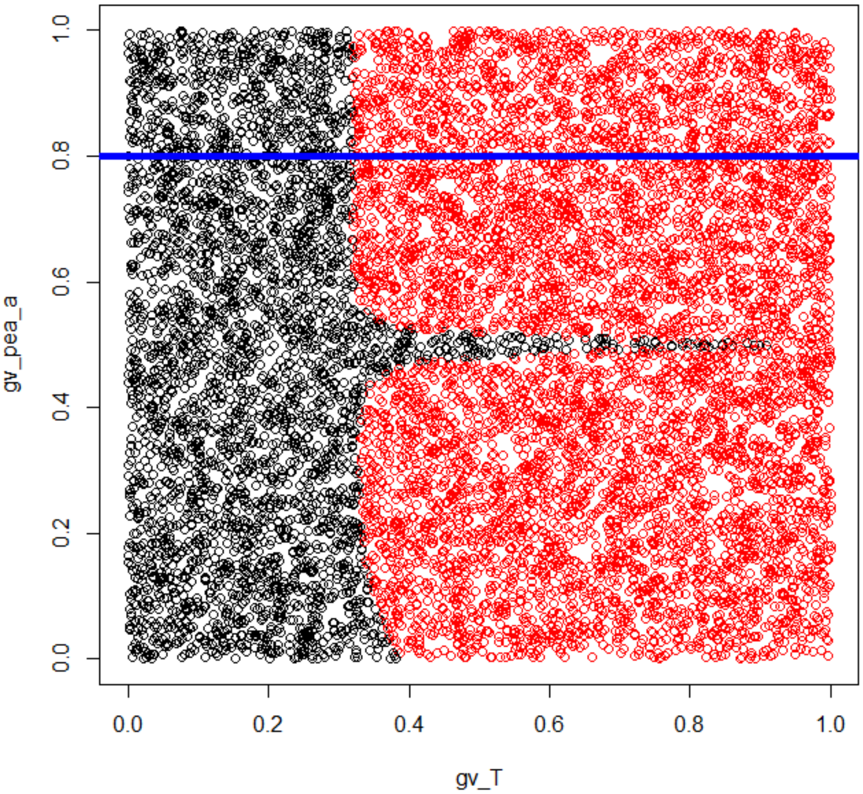

Figure 11. Suitable $T$ Interval for Table 11 ($\pi = 0.6$, $\omega = 0.7$)

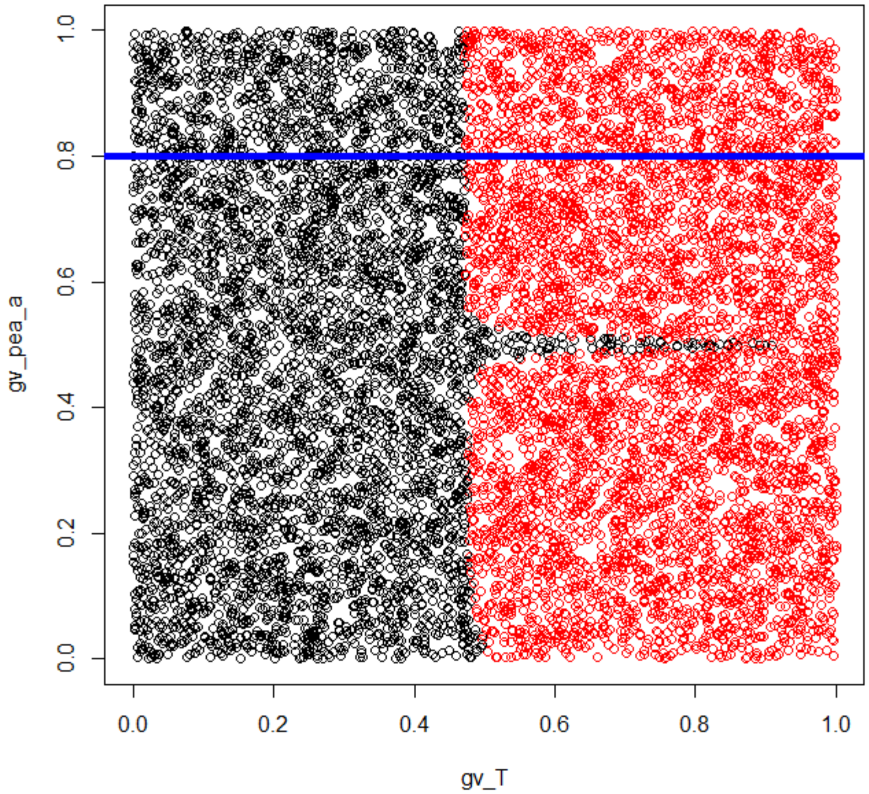

Figure 12. Suitable $T$ Interval for Table 12 ($\pi = 0.6$, $\omega = 0.9$)

# CHAPTER VI

# CONCLUSIONS AND FUTURE WORK

## 6.1 Conclusion

We propose a new binary optional RRT model to overcome shortcomings of existing models. Namely, we want to have a smaller sample size while estimating multiple parameters in surveying sensitive questions. Also, we want to avoid the unrealistic assumption of knowing the prevalence of innocuous characteristics in the two- question approach. The new model constitutes notable improvement in these matters.

We first presented the model of Sihm, Chhabra, and Gupta (2016) in Section 4.2 to show how much improvement in efficiency we could achieve by switching from the split-sample approach to the two-question approach. However, the model of Sihm, Chhabra, and Gupta (2016) has a very strong requirement for us to utilize it without difficulties. We must know the prevalence of the innocuous characters ($\pi_a$ and $\pi_b$) among the population in order to apply the method.

Assuming $\pi_a$ and $\pi_b$ are unknown, we revised the model of Sihm, Chhabra, and Gupta (2016) and came up with a more realistic version of it. But as we have to estimate four parameters instead of two, the efficiency suffered. In the end, we adopted the Warner's model into the basic framework to achieve our objectives.

The simulation study verifies that the proposed model allows us to choose a suitable value of $T$ to minimize the variance of the estimator in comparison with competing models.

The main contribution of this work is to offer a better insight in the utility of two-stage RRT models. We have discussed how the two-stage parameter $T$ (also known as the truth parameter) should be selected. Another key contribution is the introduction of the two-question approach as opposed to the traditional split-sample approach, thereby reducing the sampling cost significantly. Of course, this requires greater degree of cooperation from the respondents.

## 6.2 Future Work

We have observed that the Randomized Response Technique has been implemented mostly in surveying sensitive and private questions so far. Recently, new opportunities and applications are emerging in the field of statistical disclosure control. I would like to study more about possibilities which the RRT may open up in the context of statistical disclosure control. This is a major paradigm shift in RRT methodology where we worry about offering privacy not just to the respondent but also to data that have already been collected and are ready for public release.

Another area I would like explore will be connecting dots between model validation and actual implementation of the models. In this study, all the parameter values were freely chosen to examine how the model behaves. But in reality, the field worker wouldn't be able to know what parameter values ($\pi$ and $\omega$) he or she should take in the end.

I would also like to study the important concept of utilizing auxiliary information to improve parameter estimates.

# REFERENCES

Akers, R. L., Massey, J., Clarke, W., & Lauer, R. M. (1983). Are self-reports of adolescent deviance valid? biochemical measures, randomized response, and the bogus pipeline in smoking behavior. *Social Forces*, *62*(1), 234-251.

Brewer, K. R. W. (1981). Estimating marihuana usage using randomized response: Some paradoxical findings. *Australian Journal of Statistics*, *23*, 139-148.

Buchman, T. A., & Tracy, J. A. (1982). Obtaining responses to sensitive questions: Conventional questionnaire versus randomized response technique. *Journal of Accounting Research*, *20*, 263-271.

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the RRT and the UCT. *Sociology Methods & Research*, *40*(1–4), 169–193.

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

Elffers, H., van der Heijden, P., & Hezemans, M. (2003). Explaining regulatory non-compliance: A survey study of rule transgression for two Dutch instrumental laws, applying the randomized response method. *Journal of Quantitative Criminology*, *19*, 409-439.

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

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

Gupta, S. N. (2001). Quantifying the sensitivity level of binary response personal interview survey questions. *Journal of Combinatorics, Information & System Sciences*, *26*(1–4), 79–86.

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

Gupta, S. N., Mehta, S., Shabbir, J., & Dass, B. K. (2013). Generalized scrambling in quantitative optional randomized response models. *Communications in Statistics—Theory and Methods*, *42*(22), 4034–4042.

Gupta, S. N., & Shabbir, J. (2004). Sensitivity estimation for personal interview survey questions. *Statistica*, *64*(3), 643–653.

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

Gupta, S. N., & Thornton, B. (2002). Circumventing social desirability response bias in personal interview surveys. *American Journal of Mathematical and Management Sciences*, *22*(3-4), 369–383.

Gupta, S. N., Thornton, B., Shabbir, J., & Singhal, S. (2006). A comparison of multiplicative and additive optional RRT models. *Journal of Statistical Theory and Applications*, *64*, 226–239.

Gupta, S. N., Tuck, A., Spears Gill, T., & Crowe, M. (2013). Optional unrelated question randomized response models. *Involve: A Journal of Mathematics*, *6*(4), 483–492.

Houston, J., & Tran, A. (2001). A survey of tax evasion using the randomized response technique. *Advances in Taxation*, *13*, 69-94.

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

Krumpal, I. (2012). Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research*, *41*, 1387-1403.

Kwan, S. S. K., So, M. K. P., & Tam, K. Y. (2010). Applying the randomized response technique to elicit truthful responses to sensitive questions in is research: The case of software piracy behavior. *Information Systems Research*, *21*(4), 941–959.

Lavender, J. M., & Anderson, D. A. (2009). Effect of perceived anonymity in assessments of eating disordered behaviors and attitudes. *International Journal of Eating Disorders*, *42*(6), 546–551.

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research: 35 years of validation. *Sociological Methods and Research*, *33*(3), 319-319.

Lensvelt-Mulders, G. J. L. M., van der Heijden, P. G. M., Laudy, O., & van Gils, G. (2006). A validation of a computer-assisted randomized response survey to estimate the prevalence of fraud in social security. *Journal of the Royal Statistical Society. Series A*, *169*(2), 305–318.

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

Nayak, T. K., & Adeshiyan, S. A. (2016). On invariant post-randomization for statistical disclosure control. *International Statistics Review*, *84*(1), 26-42.

Nayak, T. K., Adeshiyan, S. A., & Zhang, C. (2016). A concise theory of randomized response techniques for privacy and confidentiality protection. In A. Chaudhuri, T. C. Christofides, & C. R. Rao (Eds.), *Handbook of statistics* (Vol. 34, p. 273-286). Elsevier.

Nayak, T. K., Zhang, C., & Adeshiyan, S. A. (2015). Emerging applications of randomized response concepts and some related issues. *Model Assisted Statistics and Applications*, *10*, 335-344.

Ostrapczuk, M., Musch, J., & Moshagen, M. (2009). A randomized-response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology*, *39*(6), 920–931.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Reiter, J., & Slavkovic, A. (2012). O privacy, where art thou?: Statistical disclosure limitation research and practice: Fascinating and growing areas of importance. *CHANCE*, *25*(1), 34-37.

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin*, *114*(2), 363–375.

Rohan, T. (2016, August 23). Antidoping agency delays publication of research. *The New York Times*, p. A1. Retrieved from `https://nyti.ms/2m2WWnB`

Schneider, A. (2003). An examination of whether incentive compensation and stock ownership affect internal auditor objectivity. *Journal of Managerial Issues*, *15*(4), 486–497.

Sihm, J. S., Chhabra, A., & Gupta, S. N. (2016). An optional unrelated question RRT model. *Involve: A Journal of Mathematics*, *9*(2), 195-209.

Sihm, J. S., & Gupta, S. (2015). A two-stage binary optional randomized response model. *Communications in Statistics - Simulation and Computation*, *44*(9), 2278-2296.

St John, F. A. V., Keane, A. M., Edwards-Jones, G. E., Jones, L., Yarnell, R. W., & Jones, J. P. G. (2011). Identifying indicators of illegal behaviour: Carnivore killing in human-managed landscapes. *Proceedings of Royal Society Biological Science*, *279*(1729), 804-812.

Striegel, H., Ulrich, R., & Simon, P. (2010). Randomized response estimates for doping and illicit drug use in elite athletes. *Drug and Alcohol Dependence*, *106*, 230-232.

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

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

Weissman, A. N., Steer, R. A., & Lipton, D. S. (1986). Estimating illicit drug use through telephone interviews and the randomized response technique. *Drug and Alcohol Dependence*, *18*, 225-233.

Wimbush, J. C., & Dalton, D. R. (1997). Base rate for employee theft: Convergence of multiple methods. *Journal of Applied Psychology*, *82*(5), 756-763.

# APPENDIX A

## R CODE FOR VARIOUS BINARY RRT COMPARISONS

```
rrt2013 <- function() {
#####################################################################
############################## 2013      ############################
#####################################################################

# HAT variables to store many trials for Empirical Values
ppai_  <-  ww_  <- numeric(gv_trials)

# Variables for Theoretical Variances
var_ppai_  <-  var_ww_  <- 0      # 10-14-2015 Unknown pi_a --> Split Sample --> var_pi_a_21 Added
                                # 10-26-2015 Unknown pi_b & pi_a --> Split Sample --> var_pi_a_23 Added



# Py1 & Py2 for 2* series
Py1   <- (1-gv_w_)*gv_pai_ + gv_w_*( gv_pea_a*gv_pai_ + (1-gv_pea_a)*gv_pai_a )
Py2   <- (1-gv_w_)*gv_pai_ + gv_w_*( gv_pea_b*gv_pai_ + (1-gv_pea_b)*gv_pai_a )

# Split Sample into n1 and n2
gv_n2    <- round( gv_sample_size/( 1+(1/gv_lambda)*sqrt(Py1*(1-Py1)/(Py2*(1-Py2) ) ) ), 0 )
gv_n1    <- (gv_sample_size - gv_n2) # 10-26-2015 Unknown pi_b --> Split Sample

tmp <- numeric(2)   #### pai_ & w_ ####

One_Trial <- function( )
{
  pea_a     <- rbinom(gv_n1 , 1, gv_pea_a) #
  w1        <- rbinom(gv_n1 , 1, gv_w_)      #
  pai1_a    <- rbinom(gv_n1 , 1, gv_pai_a)    # 10-26-2015 Unknown pi_a --> Split Sample
  pai1      <- rbinom(gv_n1 , 1, gv_pai_)

  pea_b     <- rbinom(gv_n2 , 1, gv_pea_b) #
  w2        <- rbinom(gv_n2 , 1, gv_w_)      #
  pai2_a    <- rbinom(gv_n2 , 1, gv_pai_a)    # 10-26-2015 Unknown pi_a --> Split Sample
  pai2      <- rbinom(gv_n2 , 1, gv_pai_)



  py1       <- (1-w1)*pai1 + w1*( pea_a*pai1 + (1-pea_a)*pai1_a )
  py2       <- (1-w2)*pai2 + w2*( pea_b*pai2 + (1-pea_b)*pai2_a )

  py1_hat   <- mean(py1)
  py2_hat   <- mean(py2)

  pai_hat   <- (gv_lambda*py2_hat-py1_hat)/(gv_lambda-1)
  w_hat    <- ( py1_hat - py2_hat )/
  ( (gv_pea_b - gv_pea_a)*gv_pai_a + (1-gv_pea_b)*py1_hat - (1-gv_pea_a)*py2_hat  )

  return( c(
  pai_hat , w_hat
  ) )

}



for( i in 1:gv_trials )
{

    tmp <- One_Trial( )

ppai_[i]  <- tmp[1]
ww_[i]   <- tmp[2]

}

var_ppai_    <- (1/(gv_lambda-1)^2)*(gv_lambda^2*Py2*(1-Py2)/gv_n2 + Py1*(1-Py1)/gv_n1 )

var_ww_      <- (gv_pea_b - gv_pea_a)^2*( (gv_pai_a - Py2)^2*Py1*(1-Py1)/gv_n1 +
(gv_pai_a - Py1)^2*Py2*(1-Py2)/gv_n2  )/
                 ( (gv_pea_b - gv_pea_a)*gv_pai_a + (1-gv_pea_b)*Py1 - (1-gv_pea_a)*Py2  )^4

result <- list(mean_pi_hat=mean(ppai_), var_pi_hat=var(ppai_), theoretic_var_pi_hat=var_ppai_,
mean_w_hat=mean(ww_), var_w_hat=var(ww_), theoretic_var_w_hat=var_ww_, n1=gv_n1, n2=gv_n2 )

return(result)

}



rrt2015 <- function() {
#####################################################################
############################## 2015 #################################
#####################################################################

# HAT variables to store many trials for Empirical Values
ppai_  <-  ww_  <- numeric(gv_trials)

# Variables for Theoretical Variances
var_ppai_  <-  var_ww_  <- 0      # 10-14-2015 Unknown pi_a --> Split Sample --> var_pi_a_21 Added
                                # 10-26-2015 Unknown pi_b & pi_a --> Split Sample --> var_pi_a_23 Added


Py1  <- gv_T*gv_pai_ + (1-gv_T)*( (1-gv_w_)*gv_pai_ + gv_w_*( gv_pea_a*gv_pai_ + (1-gv_pea_a)*(1-gv_pai_) ) )
Py2  <- gv_T*gv_pai_ + (1-gv_T)*( (1-gv_w_)*gv_pai_ + gv_w_*( gv_pea_b*gv_pai_ + (1-gv_pea_b)*(1-gv_pai_) ) )
```

```
# Split Sample into n1 and n2
gv_n2    <- round( gv_sample_size/( 1+(1/gv_lambda)*sqrt(Py1*(1-Py1)/(Py2*(1-Py2) ) ) ), 0 )
gv_n1    <- (gv_sample_size - gv_n2) # 10-26-2015 Unknown pi_b --> Split Sample



tmp <- numeric(2)   #### pai_ & w_ ####

One_Trial <- function( )
{
  # Define random variables for one trial
  pai1        <- rbinom(gv_n1, 1, gv_pai_)
  w1         <- rbinom(gv_n1, 1, gv_w_)

  pai2        <- rbinom(gv_n2, 1, gv_pai_)
  w2         <- rbinom(gv_n2, 1, gv_w_)

  if ( (gv_T -1)*(gv_T) == 0 ) {
  T1         <- rep(gv_T, gv_n1)
  T2         <- rep(gv_T, gv_n2)
  } else {
  T1         <- rbinom(gv_n1, 1, gv_T)
  T2         <- rbinom(gv_n2, 1, gv_T)
  }

  pea_a      <- rbinom(gv_n1, 1, gv_pea_a) # 10-14-2015 Unknown pi_a --> Split Sample
# For (pee_a_1, pee_a_2) see below
  pea_b      <- rbinom(gv_n2, 1, gv_pea_b)



  py1      <- T1*pai1 + (1-T1)*( (1-w1)*pai1 + w1*( pea_a*pai1 + (1-pea_a)*(1-pai1) ) )
  py1_hat  <- mean(py1)

  py2      <- T2*pai2 + (1-T2)*( (1-w2)*pai2 + w2*( pea_b*pai2 + (1-pea_b)*(1-pai2) ) )
  py2_hat  <- mean(py2)


  pai_hat <- (gv_lambda*py2_hat-py1_hat)/(gv_lambda-1)

  w_hat <- (py1_hat-py2_hat)/( (1-gv_T)*( (1-gv_pea_b)*(2*py1_hat-1) -
  (1- gv_pea_a)*(2*py2_hat-1)) )

  return( c(
  pai_hat , w_hat
  ) )
}
```

```
for( i in 1:gv_trials )
{

    tmp <- One_Trial( )

ppai_[i]  <- tmp[1]
ww_[i]   <- tmp[2]

}

var_ppai_ <- (1/(gv_lambda-1)^2)*(gv_lambda^2*Py2*(1-Py2)/gv_n2 + Py1*(1-Py1)/gv_n1 )

var_ww_  <- (  (gv_pea_a - gv_pea_b)^2*(  (2*Py2-1)^2*Py1*(1-Py1)/gv_n1 +
(2*Py1-1)^2*Py2*(1-Py2)/gv_n2  )  )/(  (1-gv_T)^2*( (1-gv_pea_b)*(2*Py1-1) - (1-gv_pea_a)*(2*Py2-1)  )^4  )

result <- list(mean_pi_hat=mean(ppai_), var_pi_hat=var(ppai_),
theoretic_var_pi_hat=var_ppai_, mean_w_hat=mean(ww_), var_w_hat=var(ww_),
theoretic_var_w_hat=var_ww_, n1=gv_n1, n2=gv_n2 )
return(result)

}



rrt2016 <- function(){
#####################################################################
############################## 2016 #################################
#####################################################################

# HAT variables to store many trials for Empirical Values
ppai_  <-  ww_  <- numeric(gv_trials)

# Variables for Theoretical Variances
var_ppai_  <-  var_ww_  <- 0      # 10-14-2015 Unknown pi_a --> Split Sample --> var_pi_a_21 Added
                                # 10-26-2015 Unknown pi_b & pi_a --> Split Sample --> var_pi_a_23 Added

Py1   <- gv_pea_a*gv_w_ + (1-gv_pea_a)*gv_pai_a # 10-26-2015 Unknown pi_b --> Split Sample
Py2   <- (1-gv_w_)*gv_pai_ + gv_w_*( gv_pea_b*gv_pai_ + (1-gv_pea_b)*gv_pai_b )

tmp <- numeric(2)   #### pai_ & w_ ####

One_Trial <- function( )
{

  # Define random variables for one trial
  pai_        <- rbinom(gv_sample_size, 1, gv_pai_)
  w_          <- rbinom(gv_sample_size, 1, gv_w_)


  pea_a      <- rbinom(gv_sample_size, 1, gv_pea_a)
  pea_b      <- rbinom(gv_sample_size, 1, gv_pea_b)

     pai_a    <- rbinom(gv_sample_size , 1, gv_pai_a)    #
  pai_b    <- rbinom(gv_sample_size , 1, gv_pai_b)    #


  py1       <- pea_a*w_ + (1-pea_a)*pai_a
  py1_hat   <- mean(py1)

  py2       <- (1-w_)*pai_ + w_*( pea_b*pai_ + (1-pea_b)*pai_b )
  py2_hat   <- mean(py2)


  w_hat    <- (py1_hat-(1-gv_pea_a)*gv_pai_a)/gv_pea_a
  pai_hat   <- (py2_hat-(1-gv_pea_b)*w_hat*gv_pai_b) /(1-(1-gv_pea_b)*w_hat)

  return( c(
  pai_hat , w_hat
  ) )

}




for( i in 1:gv_trials )
{

    tmp <- One_Trial( )

ppai_[i]  <- tmp[1]
ww_[i]   <- tmp[2]

}


var_ppai_    <- (1/(1-(1-gv_pea_b)*gv_w_)^2)*(Py2*(1-Py2)/gv_sample_size) +
((1-gv_pea_b)^2*(Py2-gv_pai_b)^2/(1-(1-gv_pea_b)*gv_w_)^4)*(Py1*(1-Py1)/(gv_sample_size*gv_pea_a^2))
var_ww_     <- Py1*(1-Py1)/(gv_sample_size*gv_pea_a^2)

result <- list(mean_pi_hat=mean(ppai_), var_pi_hat=var(ppai_),
theoretic_var_pi_hat=var_ppai_, mean_w_hat=mean(ww_), var_w_hat=var(ww_), theoretic_var_w_hat=var_ww_)

return(result)
}


rrt2016star <- function() {
######################################################################
############################## 2016 Star ############################
```

```
######################################################################

# HAT variables to store many trials for Empirical Values
ppai_  <-  ww_  <- numeric(gv_trials)

# Variables for Theoretical Variances
var_ppai_  <-  var_ww_  <- 0      # 10-14-2015 Unknown pi_a --> Split Sample --> var_pi_a_21 Added
                                # 10-26-2015 Unknown pi_b & pi_a --> Split Sample --> var_pi_a_23 Added

# Py1 & Py2 for 2* series
Py11   <- gv_pea_a_1*gv_w_ + (1-gv_pea_a_1)*gv_pai_a # 10-26-2015 Unknown pi_b --> Split Sample
Py12   <- gv_pea_a_2*gv_w_ + (1-gv_pea_a_2)*gv_pai_a # 10-26-2015 Unknown pi_b --> Split Sample
Py21   <- (1-gv_w_)*gv_pai_ + gv_w_*( gv_pea_b_1*gv_pai_ + (1-gv_pea_b_1)*gv_pai_b )
Py22   <- (1-gv_w_)*gv_pai_ + gv_w_*( gv_pea_b_2*gv_pai_ + (1-gv_pea_b_2)*gv_pai_b )

# Split Sample into n1 and n2
gv_n2       <- round( gv_sample_size/( 1+(1/gv_lambda_23star_b)*
sqrt(Py21*(1-Py21)/(Py22*(1-Py22) ) ) ), 0 ) # 10-26-2015 Unknown pi_b --> Split Sample
gv_n1    <- (gv_sample_size - gv_n2) # 10-26-2015 Unknown pi_b --> Split Sample

tmp <- numeric(2)    #### pai_ & w_ ####

One_Trial <- function( )
{
  pea_b_1  <- rbinom(gv_n1 , 1, gv_pea_b_1)
  pea_a_1  <- rbinom(gv_n1 , 1, gv_pea_a_1)
  w1_      <- rbinom(gv_n1 , 1, gv_w_)     #
  pai1_a    <- rbinom(gv_n1 , 1, gv_pai_a)    #
  pai1_b    <- rbinom(gv_n1 , 1, gv_pai_b)     #
  pai1      <- rbinom(gv_n1 , 1, gv_pai_)



  pea_b_2  <- rbinom(gv_n2 , 1, gv_pea_b_2) #
  pea_a_2  <- rbinom(gv_n2 , 1, gv_pea_a_2) #
  w2_       <- rbinom(gv_n2 , 1, gv_w_)     #
  pai2_a    <- rbinom(gv_n2 , 1, gv_pai_a)
  pai2_b    <- rbinom(gv_n2 , 1, gv_pai_b)
  pai2      <- rbinom(gv_n2 , 1, gv_pai_)

  # 2.3

  py21       <- (1-w1_)*pai1 + w1_*( pea_b_1*pai1 + (1-pea_b_1)*pai1_b )
  py22       <- (1-w2_)*pai2 + w2_*( pea_b_2*pai2 + (1-pea_b_2)*pai2_b )

  py21_hat   <- mean(py21)
  py22_hat   <- mean(py22)

  py11       <- pea_a_1*w1_ + (1-pea_a_1)*pai1_a
  py12       <- pea_a_2*w2_ + (1-pea_a_2)*pai2_a
  py11_hat   <- mean(py11)
  py12_hat   <- mean(py12)

  pai_hat   <- (gv_lambda_23star_b*py22_hat-py21_hat)/
  (gv_lambda_23star_b-1)
  w_hat    <- (gv_lambda_23star_a*py12_hat-py11_hat)/
  (gv_lambda_23star_a*gv_pea_a_2-gv_pea_a_1)

  return( c(
  pai_hat , w_hat
  ) )

}



for( i in 1:gv_trials )
{

    tmp <- One_Trial( )

ppai_[i]  <- tmp[1]
ww_[i]   <- tmp[2]

}

var_ppai_    <- (1/(gv_lambda_23star_b-1)^2)*(gv_lambda_23star_b^2*
Py22*(1-Py22)/gv_n2 + Py21*(1-Py21)/gv_n1 )
var_ww_     <- (1/(gv_lambda_23star_a*gv_pea_a_2-gv_pea_a_1)^2)*
(gv_lambda_23star_a^2*Py12*(1-Py12)/gv_n2 + Py11*(1-Py11)/gv_n1 )


result <- list(mean_pi_hat=mean(ppai_), var_pi_hat=var(ppai_),
theoretic_var_pi_hat=var_ppai_, mean_w_hat=mean(ww_), var_w_hat=var(ww_),
theoretic_var_w_hat=var_ww_, n1=gv_n1, n2=gv_n2 )

return(result)

}



rrt2017 <- function() {
######################################################################
############################## 2017 ##################################
######################################################################

# HAT variables to store many trials for Empirical Values
ppai_  <-  ww_  <- numeric(gv_trials)

# Variables for Theoretical Variances
var_ppai_  <-  var_ww_  <- 0      # 10-14-2015 Unknown pi_a --> Split Sample --> var_pi_a_21 Added
```

```
Py1  <- gv_pea_a*gv_w_ + (1-gv_pea_a)*(1-gv_w_)
Py2  <- gv_T*gv_pai_ + (1-gv_T)*( (1-gv_w_)*gv_pai_ +
gv_w_*( gv_pea_b*gv_pai_ + (1-gv_pea_b)*(1-gv_pai_) ) )

tmp <- numeric(2)   #### pai_ & w_ ####

One_Trial <- function( )
{
  # Define random variables for one trial
  pai_        <- rbinom(gv_sample_size, 1, gv_pai_)
  w_         <- rbinom(gv_sample_size, 1, gv_w_)

  if ( (gv_T -1)*(gv_T) == 0 ) {
  T_         <- rep(gv_T, gv_sample_size)
  } else {
  T_         <- rbinom(gv_sample_size, 1, gv_T)
  }

  pea_a      <- rbinom(gv_sample_size, 1, gv_pea_a)

  pea_b      <- rbinom(gv_sample_size, 1, gv_pea_b)

  # 3.3
  py1      <- pea_a*w_ + (1-pea_a)*(1-w_)
  py1_hat   <- mean(py1)
  py2      <- T_*pai_ + (1-T_)*( (1-w_)*pai_ + w_*
  ( pea_b*pai_ + (1-pea_b)*(1-pai_) ) )
  py2_hat  <- mean(py2)

  w_hat <- (py1_hat - (1-gv_pea_a))/(2*gv_pea_a-1)

  pai_hat <- (py2_hat-(1-gv_T)*(1-gv_pea_b)*w_hat)/
  (1-2*(1-gv_T)*(1-gv_pea_b)*w_hat)

  return( c(
  pai_hat , w_hat
  ) )
}

for( i in 1:gv_trials )
{


    tmp <- One_Trial( )

ppai_[i]  <- tmp[1]
ww_[i]   <- tmp[2]
}

var_ww_  <- Py1*(1-Py1)/(gv_sample_size*(2*gv_pea_a-1)^2)
```

```
var_ppai_ <- Py2*(1-Py2)/(gv_sample_size*(1-2*(1-gv_T)*(1-gv_pea_b)*gv_w_)^2) +
               (((1-gv_T)*(1-gv_pea_b)*(2*Py2-1))^2)*Py1*(1-Py1)/
               (gv_sample_size*(1-2*(1-gv_T)*(1-gv_pea_a)*gv_w_)^4*(2*gv_pea_a-1)^2)

result <- list(mean_pi_hat=mean(ppai_), var_pi_hat=var(ppai_),
theoretic_var_pi_hat=var_ppai_, mean_w_hat=mean(ww_), var_w_hat=var(ww_),
theoretic_var_w_hat=var_ww_ )
return(result)
}

setwd("C:/Users/SIHM/Dropbox/Research/From USB Memory/2017 Defense/R Code")

rm(list=ls())

# Setting Seed

set.seed(11)
options(scipen = 999)      ### OFF options(scipen = 999)
                           ### ON  options(scipen = 0)

gv_sample_size       <<- 1000
gv_trials            <<- 10000

gv_n1                <<-
gv_n2                <<- gv_sample_size/2    # 10-14-2015 Unknown pi_a
                                             # 10-26-2015 Unknown pi_a and pi_b for (2.3)

gv_pea_a             <<- 0.8
gv_pea_b             <<- 0.3

gv_pea_a_1           <<- 0.8 # 10-14-2015 Unknown pi_a --> Split Sample
gv_pea_a_2           <<- 0.2 # 10-14-2015 Unknown pi_a --> Split Sample

gv_pea_b_1           <<- 0.7 # 10-26-2015 Unknown pi_b --> Split Sample
gv_pea_b_2           <<- 0.4 # 10-26-2015 Unknown pi_b --> Split Sample

gv_pai_a             <<- 0.35
gv_pai_b             <<- 0.25 # 10-26-2015 (2.3a)

gv_lambda            <<- (1-gv_pea_a)/(1-gv_pea_b)  # 10-26-2015 Unknown pi_b --> Split Sample
gv_lambda_23star_a   <<- (1-gv_pea_a_1)/(1-gv_pea_a_2)  # 10-14-2015 Unknown pi_a --> Split Sample
gv_lambda_23star_b   <<- (1-gv_pea_b_1)/(1-gv_pea_b_2)  # 10-26-2015 Unknown pi_b --> Split Sample

################################################ matrix 32 by 5
################################################ run each function
gv_pai_              <<- 0.1    #  2-28-2017: 0.1, 0.2, 0.3, 0.4, 0.6, 0.7
gv_w_                <<- 0.1    #  2-28-2017: 0.1, 0.2, 0.3, 0.4, 0.6, 0.7, 0.8
        gv_T <<- 0.0    #  2-28-2017: 0.0, 0.2, 0.4, 0.8

mdata <- matrix( rep(NA, 32*5), nrow=32, ncol=5)
```

```
output <- rrt2013()
mdata[1,1]   <- mdata[9,1]   <- mdata[17,1]   <- mdata[25,1]   <- round(output$mean_pi_hat,6)
mdata[2,1]   <- mdata[10,1]   <- mdata[18,1]   <- mdata[26,1]   <- round(output$var_pi_hat,6)
mdata[3,1]   <- mdata[11,1]   <- mdata[19,1]   <- mdata[27,1]   <- round(output$theoretic_var_pi_hat,6)
mdata[4,1]   <- mdata[12,1]   <- mdata[20,1]   <- mdata[28,1]   <- round(output$mean_w_hat,6)
mdata[5,1]   <- mdata[13,1]   <- mdata[21,1]   <- mdata[29,1]   <- round(output$var_w_hat,6)
mdata[6,1]   <- mdata[14,1]   <- mdata[22,1]   <- mdata[30,1]   <- round(output$theoretic_var_w_hat,6)
mdata[7,1]   <- mdata[15,1]   <- mdata[23,1]   <- mdata[31,1]   <- round(output$n1,0)
mdata[8,1]   <- mdata[16,1]   <- mdata[24,1]   <- mdata[32,1]   <- round(output$n2,0)

output <- rrt2016()
mdata[1,3]   <- mdata[9,3]   <- mdata[17,3]   <- mdata[25,3]   <- round(output$mean_pi_hat,6)
mdata[2,3]   <- mdata[10,3]   <- mdata[18,3]   <- mdata[26,3]   <- round(output$var_pi_hat,6)
mdata[3,3]   <- mdata[11,3]   <- mdata[19,3]   <- mdata[27,3]   <- round(output$theoretic_var_pi_hat,6)
mdata[4,3]   <- mdata[12,3]   <- mdata[20,3]   <- mdata[28,3]   <- round(output$mean_w_hat,6)
mdata[5,3]   <- mdata[13,3]   <- mdata[21,3]   <- mdata[29,3]   <- round(output$var_w_hat,6)
mdata[6,3]   <- mdata[14,3]   <- mdata[22,3]   <- mdata[30,3]   <- round(output$theoretic_var_w_hat,6)

output <- rrt2016star()
mdata[1,4]   <- mdata[9,4]   <- mdata[17,4]   <- mdata[25,4]   <- round(output$mean_pi_hat,6)
mdata[2,4]   <- mdata[10,4]   <- mdata[18,4]   <- mdata[26,4]   <- round(output$var_pi_hat,6)
mdata[3,4]   <- mdata[11,4]   <- mdata[19,4]   <- mdata[27,4]   <- round(output$theoretic_var_pi_hat,6)
mdata[4,4]   <- mdata[12,4]   <- mdata[20,4]   <- mdata[28,4]   <- round(output$mean_w_hat,6)
mdata[5,4]   <- mdata[13,4]   <- mdata[21,4]   <- mdata[29,4]   <- round(output$var_w_hat,6)
mdata[6,4]   <- mdata[14,4]   <- mdata[22,4]   <- mdata[30,4]   <- round(output$theoretic_var_w_hat,6)
mdata[7,4]   <- mdata[15,4]   <- mdata[23,4]   <- mdata[31,4]   <- round(output$n1,0)
mdata[8,4]   <- mdata[16,4]   <- mdata[24,4]   <- mdata[32,4]   <- round(output$n2,0)

        gv_T <<- 0.0    #  2-28-2017: 0.0, 0.2, 0.4, 0.8
output <- rrt2015()
mdata[1,2]   <- round(output$mean_pi_hat,6)
mdata[2,2]   <- round(output$var_pi_hat,6)
mdata[3,2]   <- round(output$theoretic_var_pi_hat,6)
mdata[4,2]   <- round(output$mean_w_hat,6)
mdata[5,2]   <- round(output$var_w_hat,6)
mdata[6,2]   <- round(output$theoretic_var_w_hat,6)
mdata[7,2]   <- round(output$n1,0)
mdata[8,2]   <- round(output$n2,0)

output <- rrt2017()
mdata[1,5]   <- round(output$mean_pi_hat,6)
mdata[2,5]   <- round(output$var_pi_hat,6)
mdata[3,5]   <- round(output$theoretic_var_pi_hat,6)
mdata[4,5]   <- round(output$mean_w_hat,6)
mdata[5,5]   <- round(output$var_w_hat,6)
mdata[6,5]   <- round(output$theoretic_var_w_hat,6)
gv_T <<- 0.2    #  2-28-2017: 0.0, 0.2, 0.4, 0.8
output <- rrt2015()
mdata[9,2]   <- round(output$mean_pi_hat,6)
mdata[10,2]   <- round(output$var_pi_hat,6)
```

```
mdata[11,2]   <- round(output$theoretic_var_pi_hat,6)
mdata[12,2]   <- round(output$mean_w_hat,6)
mdata[13,2]   <- round(output$var_w_hat,6)
mdata[14,2]   <- round(output$theoretic_var_w_hat,6)
mdata[15,2]   <- round(output$n1,0)
mdata[16,2]   <- round(output$n2,0)

output <- rrt2017()
mdata[9,5]   <- round(output$mean_pi_hat,6)
mdata[10,5]   <- round(output$var_pi_hat,6)
mdata[11,5]   <- round(output$theoretic_var_pi_hat,6)
mdata[12,5]   <- round(output$mean_w_hat,6)
mdata[13,5]   <- round(output$var_w_hat,6)
mdata[14,5]   <- round(output$theoretic_var_w_hat,6)


        gv_T <<- 0.4    #  2-28-2017: 0.0, 0.2, 0.4, 0.8
output <- rrt2015()
mdata[17,2]   <- round(output$mean_pi_hat,6)
mdata[18,2]   <- round(output$var_pi_hat,6)
mdata[19,2]   <- round(output$theoretic_var_pi_hat,6)
mdata[20,2]   <- round(output$mean_w_hat,6)
mdata[21,2]   <- round(output$var_w_hat,6)
mdata[22,2]   <- round(output$theoretic_var_w_hat,6)
mdata[23,2]   <- round(output$n1,0)
mdata[24,2]   <- round(output$n2,0)

output <- rrt2017()
mdata[17,5]   <- round(output$mean_pi_hat,6)
mdata[18,5]   <- round(output$var_pi_hat,6)
mdata[19,5]   <- round(output$theoretic_var_pi_hat,6)
mdata[20,5]   <- round(output$mean_w_hat,6)
mdata[21,5]   <- round(output$var_w_hat,6)
mdata[22,5]   <- round(output$theoretic_var_w_hat,6)



        gv_T <<- 0.8    #  2-28-2017: 0.0, 0.2, 0.4, 0.8
output <- rrt2015()
mdata[25,2]   <- round(output$mean_pi_hat,6)
mdata[26,2]   <- round(output$var_pi_hat,6)
mdata[27,2]   <- round(output$theoretic_var_pi_hat,6)
mdata[28,2]   <- round(output$mean_w_hat,6)
mdata[29,2]   <- round(output$var_w_hat,6)
mdata[30,2]   <- round(output$theoretic_var_w_hat,6)
mdata[31,2]   <- round(output$n1,0)
mdata[32,2]   <- round(output$n2,0)

output <- rrt2017()
mdata[25,5]   <- round(output$mean_pi_hat,6)
```

```
mdata[26,5]   <- round(output$var_pi_hat,6)
mdata[27,5]   <- round(output$theoretic_var_pi_hat,6)
mdata[28,5]   <- round(output$mean_w_hat,6)
mdata[29,5]   <- round(output$var_w_hat,6)
mdata[30,5]   <- round(output$theoretic_var_w_hat,6)

latextable(mdata, scientific=0, digits=6)
```

# APPENDIX B

# R CODE FOR VARIANCE COMPARISON AND FINDING SUITABLE TRUTH PARAMETER FOR THE TWO-STAGE RRT

```
setwd("C:/Users/SIHM/Dropbox/Research/From USB Memory/2017 Defense/R Code")

rm(list=ls())

# Setting Seed

set.seed(11)
options(scipen = 999)      ### OFF options(scipen = 999)
                           ### ON  options(scipen = 0)

gv_sample_size       <<- 1000
gv_trials            <<- 10000

gv_n1                <<-
gv_n2                <<- gv_sample_size/2    # 10-14-2015 Unknown pi_a
                                             # 10-26-2015 Unknown pi_a and pi_b for (2.3)

### gv_pea_a             <<- 0.8 # Selected For Simulation 03-08-2017
gv_pea_a             <-  runif(gv_trials, 0, 1)
gv_pea_b             <<- 0.3

gv_pea_a_1           <<- 0.8 # 10-14-2015 Unknown pi_a --> Split Sample
gv_pea_a_2           <<- 0.2 # 10-14-2015 Unknown pi_a --> Split Sample

gv_pea_b_1           <<- 0.7 # 10-26-2015 Unknown pi_b --> Split Sample
gv_pea_b_2           <<- 0.4 # 10-26-2015 Unknown pi_b --> Split Sample

gv_pai_a             <<- 0.35
gv_pai_b             <<- 0.25 # 10-26-2015 (2.3a)

gv_lambda            <<- (1-gv_pea_a)/(1-gv_pea_b)  # 10-26-2015 Unknown pi_b --> Split Sample
gv_lambda_23star_a   <<- (1-gv_pea_a_1)/(1-gv_pea_a_2)  # 10-14-2015 Unknown pi_a --> Split Sample
gv_lambda_23star_b   <<- (1-gv_pea_b_1)/(1-gv_pea_b_2)  # 10-26-2015 Unknown pi_b --> Split Sample

gv_pai_              <<- 0.1    #  2-28-2017: 0.1, 0.2, 0.3, 0.4, 0.6, 0.7
gv_w_                <<- 0.1    #  2-28-2017: 0.1, 0.2, 0.3, 0.4, 0.6, 0.7, 0.8
#         gv_T <<- 0.0    #  2-28-2017: 0.0, 0.2, 0.4, 0.8  # Selected For Simulation 03-08-2017
gv_T                 <-  runif(gv_trials, 0, 1)

################################# 2016 Star #################################
# Py1 & Py2 for 2* series
Py11   <- gv_pea_a_1*gv_w_ + (1-gv_pea_a_1)*gv_pai_a # 10-26-2015 Unknown pi_b --> Split Sample
```

```
Py12   <- gv_pea_a_2*gv_w_ + (1-gv_pea_a_2)*gv_pai_a # 10-26-2015 Unknown pi_b --> Split Sample
Py21   <- (1-gv_w_)*gv_pai_ + gv_w_*( gv_pea_b_1*gv_pai_ + (1-gv_pea_b_1)*gv_pai_b )
Py22   <- (1-gv_w_)*gv_pai_ + gv_w_*( gv_pea_b_2*gv_pai_ + (1-gv_pea_b_2)*gv_pai_b )

# Split Sample into n1 and n2
gv_n2    <- round( gv_sample_size/( 1+(1/gv_lambda_23star_b)*sqrt(Py21*(1-Py21)/(Py22*(1-Py22) ) ) ), 0 )
# 10-26-2015 Unknown pi_b --> Split Sample
gv_n1    <- (gv_sample_size - gv_n2) # 10-26-2015 Unknown pi_b --> Split Sample

var_ppai_2016_star   <- (1/(gv_lambda_23star_b-1)^2)*(gv_lambda_23star_b^2*Py22*(1-Py22)/gv_n2 + Py21*(1-Py21)/gv_n1 )

#################################### 2017 #####################################

Py1  <- gv_pea_a*gv_w_ + (1-gv_pea_a)*(1-gv_w_)
Py2  <- gv_T*gv_pai_ + (1-gv_T)*( (1-gv_w_)*gv_pai_ + gv_w_*( gv_pea_b*gv_pai_ + (1-gv_pea_b)*(1-gv_pai_) ) )

var_ppai_2017 <- Py2*(1-Py2)/(gv_sample_size*(1-2*(1-gv_T)*(1-gv_pea_b)*gv_w_)^2) +
               (((1-gv_T)*(1-gv_pea_b)*(2*Py2-1))^2)*Py1*(1-Py1)/(gv_sample_size*(1-2*(1-gv_T)*(1-gv_pea_a)*gv_w_)^4*
               (2*gv_pea_a-1)^2)

################################# Print Out ###################################

var_ppai_2016_star; min(var_ppai_2017); max(var_ppai_2017); min(var_ppai_2016_star/var_ppai_2017);
max(var_ppai_2016_star/var_ppai_2017)

################################# Graph ###################################

plot(gv_T, gv_pea_a, col=( (var_ppai_2016_star >= var_ppai_2017) + 1) )
```