NEW APPROACHES TO RANDOMIZED RESPONSE TECHNIQUE

By

JONG-MIN KIM

Bachelor of Science in Mathematics Education
Chongju University
Chongju, South Korea
1994

Master of Science in Mathematics
Chung-Ang University
Seoul, South Korea
1996

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2002

NEW APPROACHES TO RANDOMIZED RESPONSE

TECHNIQUE

Thesis Approved:

____
Thesis Adviser

____

____

____

____
Dean of the Graduate College

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my major advisor, Dr. William D. Warde, for providing support, guidance, and assistance during the course of this study. I would also like to thank the members of my graduate committee, Dr. Mark E. Payton, Dr. Brenda J. Masters, and Dr. Dennis Bertholf, for reviewing my work and providing many helpful comments.

I would like to thank Stillwater people, including faculty, staff, and friends, for all of their encouragement and support while I have been in Stillwater.

I would like to thank my father, Chang-Bae Kim, my mother, Young-Soon Jun, my brothers, my sisters, and my relatives for their endless love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1. Introduction to an Alternative Survey Technique

Survey estimates are affected by two main sources of error. The first type of error is sampling error that results from taking a sample instead of enumerating the whole population. The second type of error is non-sampling error that cannot be attributed to sample-to-sample variability. Non-sampling error has two different errors which is random error and nonrandom error. Random error, which results from reducing the reliability of measurements, can be minimized over repeated measurements. But nonrandom error that is bias in survey data is difficult to cancel out over repeated measurements. Deming (1960) and Cochran (1977) have discussed the sources of non-sampling error and its effects on sampling estimates. The main sources of non-sampling error in any survey are non-response bias and response bias. Non-response bias arises from subjects' refusal to respond and response bias arises from giving incorrect responses. When open or direct surveys are about sensitive matters, for example, gambling habits, addiction to drugs and other intoxicants, alcoholism, proneness to tax evasion, induced abortions, drunken driving, history of past involvement in crimes, and homosexuality, non-response bias and response bias become serious because people usually do not wish to give correct information. A survey technique that encourages truthful answers but makes people comfortable was necessary instead of open or direct surveys. Warner (1965) developed such an alternative survey technique that is called to randomized response technique. Warner's randomized response survey technique is designed to eliminate evasive answer bias and keep Respondents' confidentiality. Since Warner presented the randomized response technique, many variants of the Warner model have been presented. One of the variants is the unrelated randomized response model presented by Greenberg et al. (1969).

In the article by Campbell and Joiner (1973), Tom Hettmansperger, a statistics professor, surveyed his class to estimate the proportion of regular "pot" smokers on the campus. He applied the unrelated question randomized response model to students in his class. The sensitive question was, "Do you smoke pot at least once a week?" The unrelated question was, "Is the last digit of your student ID number odd?" After finished the survey, he estimated that 41% of his students used pot at least once a week. The students verified voluntarily the proportion of regular "pot" smokers in the classroom.

The students found that 38% of the students in the class were regular pot smokers. It turned out that the estimated proportion is quite close to the true proportion, so this is a good example to show that the randomized response survey technique works well.

Validation checks for randomized response technique have also been attempted by Abernathy et al. (1970), Bradburn and Sudman (1979), Tracy and Fox (1981), Danermark and Swensson (1987), Duffy and Waterton (1988) and Kerkvliet (1994). These researchers did the comparison of RR interview and direct interview based on statistical measures of efficiency and respondents' protection. Tracy and Fox (1981) conducted a field-validation of a quantitative randomized response model. They compared self-reports of arrests obtained in a direct question condition with estimates obtained from randomized response. True scores regarding the number of official arrests from criminal history records were available as a validation criterion. Table 1.1 show that constant bias and especially systematic bias were much more substantial in the direct question condition than in randomized response.

TABLE 1.1

Two or More Arrests Case in the TRACYAND FOX (1981)

Two or more Arrests	Sample size (n)	Mean Reported Arrests	Mean Official Arrests
RR technique	84	2.7286	3.2024
Direct question	40	1.5500	3.3500

1.2. Objectives and Brief Summary of the Study

It is the objective of this dissertation to develop new randomized response models that increase the cooperation of the respondents, decrease the variances of the randomized response estimators and investigate the properties of the randomized response models.

In Chapter II, the literature of randomized response models is briefly reviewed in two different sections. One section is in terms of literature review on qualitative randomized response models and the other section is in terms of literature review on quantitative randomized response models.

In Chapter III, a new stratified randomized response model that is more efficient than the Hong et al. (1994) stratified randomized response model is presented. In this research, a drawback of the Hong et al. model under their proportional sampling assumption is pointed out. The proposed stratified randomized response model has an optimal allocation and large gain in precision. Hence, it is shown that the estimator based on the proposed method is more efficient than the Warner (1965), the Mangat and Singh (1990) and the Mangat (1994) estimators under the conditions presented in both the case of completely truthful reporting and that of not completely truthful reporting by respondents.

In Chapter IV, a mixed randomized response model is introduced. It consists of Warner's model and Simmons' model. The proposed model is a variation of Lanke's (1975) idea. Mangat et al. (1997) and Singh et al. (2000) found a privacy problem and presented several strategies as an alternative one for the Moors' model, but their models may lose a large portion of information and require a high cost to obtain confidentiality for the respondents.

Our proposed model has the advantage of simplicity over the previous models while keeping the confidentiality of the interviewee. Furthermore, the mixed model will be extended to a stratified mixed RR model.

In Chapter V, a new quantitative randomized response technique is presented. The proposed technique will use a Hopkins' randomizing device to derive a multinomial distribution for sensitive categories. After obtaining the observed estimates for sensitive category proportions which also include the random responses from the Hopkins' randomizing device, we derive the true estimates of the proportions for the sensitive categories. For contingency tables, a Pearson product-moment correlation between two different sensitive questions will be derived.

CHAPTER II

REVIEW OF THE RANDOMIZED RESPONSE MODEL

2.1. Literature Review on Qualitative Randomized Response Model

In initiating the work on Randomized Response Technique, Warner (1965) presented a two related question model for estimating the population proportion of people who possess a sensitive trait in a given population. To apply the Warner model, a simple random sample of n people will be drawn with replacement from the population and each person will be interviewed. Before the interviews, each interviewer is furnished with an identical spinner (randomization device) which points to Statement 1 with probability P, and to Statement 2 with probability $1-P$. Without reporting the outcome of the spinner to the interviewer, the interviewee answers one of the following statements:

Statement 1: I belong to the sensitive trait group.

Statement 2: I do not belong to the sensitive trait group.

depending on the outcome directed by the randomization device. Warner equated the proportion of respondents who answer "Yes" to Statement 1 or to Statement 2:

$$X = P\pi_s + (1-P)(1-\pi_s) \tag{2.1.1}$$

where X is the proportion of "Yes" responses, π_s is the proportion of people with the sensitive trait. Under the assumption that the total number of "Yes" responses is known from the sample and $P(\neq 0.5)$ is set by a researcher, the maximum likelihood estimator of π_s is

$$\hat{\pi}_w = \frac{X-(1-P)}{2P-1}. \tag{2.1.2}$$

Warner (1965) has shown that $\hat{\pi}_W$ is an unbiased estimator of π_S and the variance of $\hat{\pi}_W$ is

$$Var(\hat{\pi}_W) = \frac{\pi_S(1-\pi_S)}{n} + \frac{P(1-P)}{n(2P-1)^2}. \tag{2.1.3}$$

where n is the total number of units in the sample.

Greenberg et al. (1969) developed the theoretical framework for the unrelated question RR model even if Horvitz et al. (1967) developed this model. Contrary to the Warner model, the unrelated question RR model has one question that asks about very sensitive trait and the other question ask about an innocuous or non-sensitive trait.

Greenberg et al. (1969) proposed two models; the case of unknown π_I, the proportion of people with an innocuous trait, and the case of known π_I. Let's explain the unrelated question RR model for the case of unknown π_I. Using simple random sampling with replacement, two samples with sizes n_1 and n_2 are independently drawn from the population. Each interviewee in the i sample is required to use a randomization device with two outcomes with preassigned probabilities, P_i and $1-P_i$, for $i = 1, 2$. Without reporting the outcome of the spinner to the interviewer, the interviewee answers "Yes" or "No" to one of the following statements:

Statement A: I belong to the sensitive trait group.

Statement B: I belong to the innocuous trait group.

depending on the outcome from the randomization device.

Figure 1.1. Warner's Randomizing Device

The proportion of respondents who answer “Yes” to Statement A or to Statement B as follows:

$$Y_i = P_i \pi_S + (1 - P_i)\pi_I \qquad \text{for } i = 1, 2. \tag{2.1.4}$$

where Y_i is the proportion of “Yes” responses, π_S is the proportion of people with the sensitive trait. Under the assumption that the total number of “Yes” responses is known from the sample and P_i is set by the researcher, the unbiased estimator for π_S is

$$\hat{\pi}_U = \frac{(1 - P_2)\hat{Y}_1 - (1 - P_1)\hat{Y}_2}{P_1 - P_2}. \tag{2.1.5}$$

Since $Var(\hat{Y}_i) = Y_i(1 - Y_i)/n_i$ and $\hat{Y}_1$ and $\hat{Y}_2$ are independent, they derived the variance of $\hat{\pi}_U$:

$$Var(\hat{\pi}_U) = \frac{1}{(P_1 - P_2)^2}\left[\frac{(1 - P_2)^2 Y_1(1 - Y_1)}{n_1} + \frac{(1 - P_1)^2 Y_2(1 - Y_2)}{n_2}\right]. \tag{2.1.6}$$

Consider the simple case when π_I, the true proportion in group G in the population, is known. A simple random sample with replacement of size n is drawn from the population and each interviewee is asked to report only “yes” or “no” regarding belonging to the sensitive trait group (chosen with probability P) or to the innocuous trait group (chosen with probability $1 - P$). The probability of a “yes” response is

$$Y = P\pi_S + (1 - P)\pi_I. \tag{2.1.7}$$

The unbiased estimator for π_S is

$$\hat{\pi}_{UK} = \frac{\hat{Y} - (1 - P)\pi_I}{P}. \tag{2.1.8}$$

The variance of $\hat{\pi}_{UK}$ is

$$Var(\hat{\pi}_{UK}) = \frac{Y(1-Y)}{nP^2}. \tag{2.1.9}$$

Moors (1971) presented a variation of unrelated question randomized response model which has the advantage that, even when questions having known distributions are not available, the simplicity of the known π_I model can be achieved if one of the samples in unrelated question randomized response model were used exclusively to estimate an unknown π_I.

The Moors model has a characteristic that one of the two independent samples would be used to estimate the proportion of people who possess the innocuous trait by way of the direct question. Setting $P_2 = 0$ in (2.1.5) gives the Moors (1971) model. The unbiased estimator for π_S is

$$\hat{\pi}_{UM} = \frac{\hat{Y}_1 - (1-P_1)\hat{Y}_2}{P_1}. \tag{2.1.10}$$

Setting $P_2 = 0$ in (2.1.6), the variance of $\hat{\pi}_{UM}$ is

$$Var(\hat{\pi}_{UM}) = \frac{1}{P_1^2}\left[\frac{Y_1(1-Y_1)}{n_1} + \frac{(1-P_1)^2 Y_2(1-Y_2)}{n_2}\right]. \tag{2.1.11}$$

Mangat and Singh (1990) proposed a two-stage randomized response model that is a variation of the Warner model. In this model, each interviewee in the simple random sample with replacement of n respondents is provided with two random devices. The random device R_1 consists of two statements. The one statement is that "I belong to the sensitive trait group" (with probability M), and the other statement is that "Go to random device R_2 "(with probability $1-M$). The random device R_2 also consists of two statements which are "I belong to the sensitive group" and "I do not belong to the sensitive group" with known probabilities P and $1-P$, is the same as used by Warner (1965). They derived that the proportion of respondents who answer "Yes" for the sensitive question and the negative of the sensitive question is

$$\theta = M\pi_S + (1-M)[P\pi_S + (1-P)(1-\pi_S)] \tag{2.1.12}$$

where θ is the proportion of "Yes" responses.

It is assumed that the total number of "Yes" responses is known from the sample and M and $P\ (\neq 0.5)$ are set by the researcher. The maximum likelihood estimator is

$$\hat{\pi}_{ms} = \frac{\hat{\theta} - (1-M)(1-P)}{2P - 1 + 2M(1-P)}. \tag{2.1.13}$$

Mangat and Singh (1990) showed that the variance of an unbiased estimator $\hat{\pi}_{ms}$ is

$$Var(\hat{\pi}_{ms}) = \frac{\pi_S(1-\pi_S)}{n} + \frac{(1-M)(1-P)[1-(1-M)(1-P))]}{n[2P-1+2M(1-P)]^2} \tag{2.1.14}$$

and the mean square error of $\hat{\pi}_{ms}$ in the case of less than completely truthful "Yes" answer to the sensitive statement and to the negative form of the statement is

$$MSE(\hat{\pi}_{ms}) = \frac{\pi_S T_r(1-\pi_S T_r)}{n} + \frac{(1-M)(1-P)[1-(1-M)(1-P)]}{n[2P-1+2M(1-P)]^2} + [\pi_S(T_r - 1)]^2$$

$$+\pi_S M(T-T_r)[1+\pi_S(n-1)\{M(T-T_r)+4MT_r(1-P)+2T_r(2P-1)\}$$

$$-2(1-M)(1-P)-2\pi_S n\{2M(1-P)+2P-1\}][n\{2P-1+2M(1-P)\}^2]^{-1} \tag{2.1.15}$$

where T and T_r are the probabilities that a respondent with the sensitive trait will report truthfully at the first stage and second stage.

Mangat (1994) proposed another RR model which has the benefit of simplicity over that of Mangat and Singh (1990).

The probability of a "Yes" response for this model is given by

$$Y_M = \pi_S + (1-P)(1-\pi_S) \tag{2.1.16}$$

where Y_M is the proportion of "Yes" responses and P is the probability of selecting the sensitive question.

Mangat (1994) showed that the variance of an unbiased estimator $\hat{\pi}_m$ is

$$Var(\hat{\pi}_m) = \frac{\pi_S(1-\pi_S)}{n} + \frac{(1-\pi_S)(1-P)}{nP} \tag{2.1.17}$$

and the mean square error of $\hat{\pi}_m$ in the case of less than completely truthful "Yes" answer to the sensitive statement and to the negative form of the statement is

$$MSE(\hat{\pi}_m) = \frac{\pi_S T_r(1-\pi_S T_r)}{nP^2} + \frac{(1-\pi_S)(1-P)[1-(1-\pi_S)(1-P)-2\pi_S T_r]}{nP^2} + \left[\frac{\pi_S(T_r-1)}{P}\right]^2 \tag{2.1.18}$$

where T_r is the probability that a respondent with the sensitive trait will report truthfully.

This section briefly reviewed the literature of qualitative randomized response techniques including the original Warner model and other randomized response models.

2.2. Literature Review on Quantitative Randomized Response Model

Greenberg et al. (1971) have proposed the unrelated question randomized response design for estimating the mean and the variance of the distribution of a quantitative variable. The RR technique is similar to the unrelated question RR technique of Horvitz et al. (1967) in terms of a survey procedure that a respondent could be asked one of two questions depending on the outcome of some randomization device. For example, an interviewee to perform a randomization device with two outcomes with pre-assigned probabilities P and $1-P$ will answer one of the following questions:

Sensitive question: How many abortions have you had during your lifetime?

Non-sensitive question: How many magazines do you subscribe to?

Two independent samples of sizes n_1 and n_2 are employed. Unbiased estimators for the means of the sensitive and non-sensitive distributions, μ_A and μ_B respectively, are

$$\hat{\mu}_A = \frac{(1-P_2)\overline{T}_1 - (1-P_1)\overline{T}_2}{P_1 - P_2} \tag{2.2.1}$$

$$\hat{\mu}_B = \frac{P_2\overline{T}_1 - P_1\overline{T}_2}{P_2 - P_1} \tag{2.2.2}$$

where $\overline{T}_1$ and $\overline{T}_2$ are sample means computed from the responses in the two samples; and P_j is the selection probability for the sensitive question in the j th sample ($P_1 \neq P_2$).

The variance of this estimate is given by

$$Var(\hat{\mu}_A) = \frac{(1-P_2)^2 Var(\overline{T}_1) + (1-P_1)^2 Var(\overline{T}_2)}{(P_2 - P_1)^2} \tag{2.2.3}$$

where $Var(\overline{T}_j) = \frac{1}{n_j}\left[\sigma_B^2 + P_j(\sigma_A^2 - \sigma_B^2) + P_j(1-P_j)(\mu_A - \mu_B)^2\right].$

Only one sample is required if μ_B and σ_B are known in advance. Furthermore, it can be derived the result as a substantial reduction in the variance of $\hat{\mu}_A$ by an empirical investigation from the qualitative unrelated question randomized response technique.

Equations (2.2.1) and (2.2.3) simplify to

$$\hat{\mu}_A = \frac{\bar{T} - (1-P)\mu_B}{P} \tag{2.2.4}$$

and

$$Var(\hat{\mu}_A) = \frac{Var(\bar{T})}{P^2}. \tag{2.2.5}$$

Eriksson (1973) has presented a discrete quantitative RR technique which modified the quantitative unrelated question RR technique by Greenberg et al. (1971). Eriksson used a deck of cards which consist of two different types of cards. The first type card is a red card and the second type of cards is a card with a designated number (B_i). If a respondent possesses a red card then she or he should answer the sensitive question (A). Otherwise, if a respondent possesses the second type of cards then she or he should say the designated number (B_i). The randomization device with two types of cards with preassigned probabilities P and $1-P$ will give each respondent one of the following statements:

Statement 1: Give a truthful answer for A.

Statement 2: Just say the designated number B_i.

The proportion of cards with designated number B_i is p_i such that $1-P=\sum_1^k p_i$.

Mean and variance for a designated number B_i are as follows:

$$\mu_B = \sum_{i=1}^{k} B_i \frac{p_i}{1-P} \tag{2.2.6}$$

$$\sigma_B^2 = \sum_{i=1}^{k} (B_i - \mu_B)^2 \frac{p_i}{1-P} \tag{2.2.7}$$

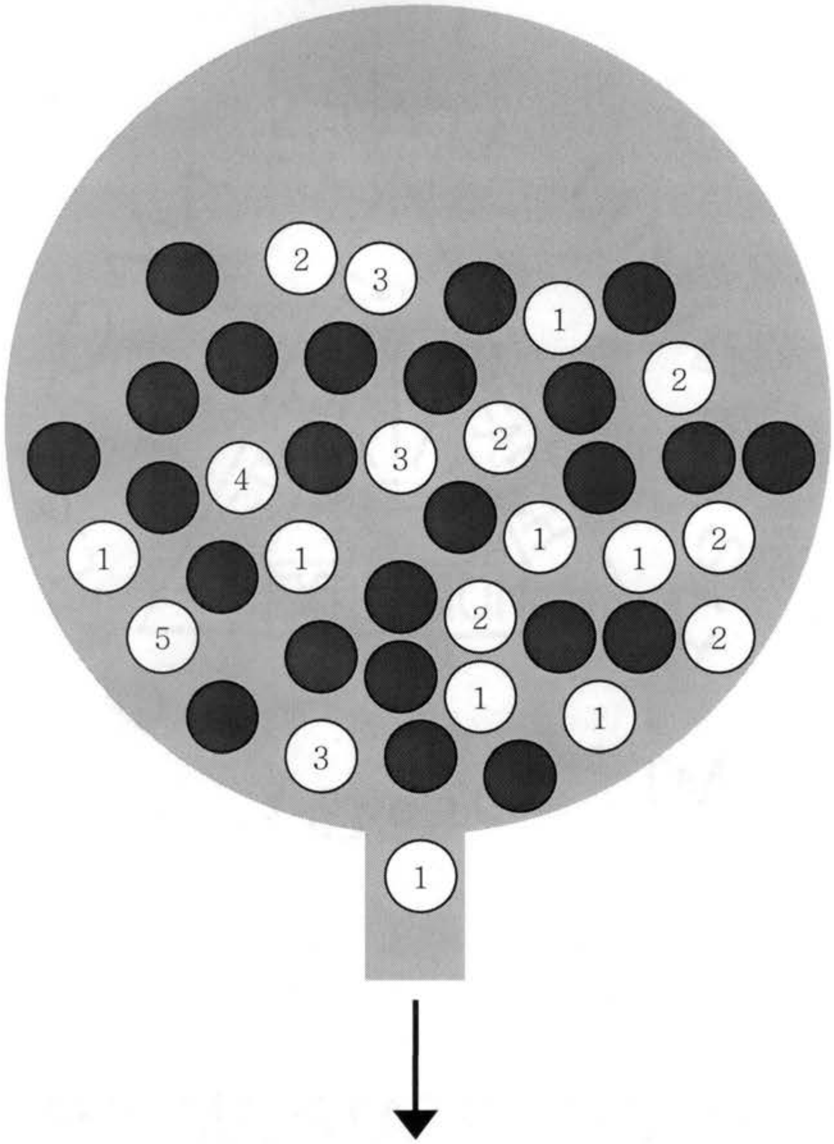

Figure 2.1. Hopkin's Randomizing Device

Suppose that there is an arbitrary sample of n respondents in the survey. The unbiased estimator of μ_A at randomized response is given by

$$\hat{\mu}_A = \frac{\overline{T}_i - (1-P)\mu_B}{P} \tag{2.2.8}$$

where $\overline{T}_i$ is the mean response of all n respondents. The variance of $\hat{\mu}_A$ is given by

$$Var(\hat{\mu}_A) = \frac{\sigma_A^2}{n} + \frac{1-P}{nkP}\left[\frac{\sigma_B^2}{P} + \sigma_A^2 + (\mu_A - \mu_B)^2\right]. \tag{2.2.9}$$

Instead of using a card type device, Liu and Chow (1976) proposed the Hopkins' randomizing device which consists of a jar containing red color balls and green color balls. Each of the green balls has a discrete number mark, such as $0,1,2,\cdots,m.$

We denote g_i to be the number of green balls with i figure and r denote the number of red balls. So the total number of balls in the device is $r + \sum_{i=0}^{m} g_i = r + g$. The proportion of red to green balls, and of green balls with different number, will be predefined. A respondent is asked to turn the device upside down, shake the device thoroughly, and turn it right side up to allow one of the balls to appear in the window of the device. If a respondent possesses a red ball then she or he should answer a sensitive question (A). Otherwise, if a respondent possesses a green ball with i figure then she or he should say a designated number i . To protect the respondents' privacy, interviewers stand on the opposite side of the window of the device. Therefore interviewers do not know whether the respondents have been asked to respond to the sensitive question or whether the respondents are responding with the number on a green ball. Let π_i represent the true proportion of respondents belonging to a sensitive trait category i .

Liu and Chow (1976) derived the unbiased estimator of π_i like this:

$$\hat{\pi}_i = \frac{(r+g_i)\hat{P}_i}{r} - \frac{g_i}{r} \tag{2.2.10}$$

where $\hat{P}_i$ is estimate of the probability that a respondent randomly selected from a population will give an answer i. Suppose that there is an arbitrary sample of n respondents in the survey. Then the estimate of variance and covariance for $\hat{\pi}_i$ are

$$v(\hat{\pi}_i) = \frac{(r+g_i)^2 \hat{P}_i(1-\hat{P}_i)}{r^2 n} \tag{2.2.11}$$

and

$$C\hat{o}v(\hat{\pi}_i, \hat{\pi}_j) = -\frac{(r+g_i)^2 \hat{P}_i \hat{P}_j}{r^2 n} \tag{2.2.12}$$

respectively.

TABLE 2.1

Randomized Response Models Introduced in the Literature Review

Authors / Year	*Qualitative/Quantitative*	*Characteristic*
Warner, S. L. (1965)	Qualitative RR Model	Original RR Model
Greenberg, B.G. et. Al. (1969)	Qualitative RR Model	Unrelated Question
Greenberg, B.G. et. Al. (1969)	Quantitative RR Model	
Eriksson, S.A. (1973)	Quantitative RR Model	Card with a discrete figure
Liu, P.T., and Chow, L.P. (1976)	Quantitative RR Model	Hopkins' randomizing device
Mangat, N.S. and Singh, R. (1990)	Qualitative RR Model	Two-Stage RR Model
Mangat, N.S. (1994)	Qualitative RR Model	

CHAPTER III

A STRATIFIED WARNER'S RANDOMIZED RESPONSE MODEL

3.1. Introduction

Warner (1965) did the pioneering work of a randomized response (RR) technique which minimizes underreporting of data relative to socially undesirable or incriminating behavior. Researchers such as Horvitz et al. (1967), Greenberg et al. (1969), Chaudhuri and Mukerjee (1988), Kuk (1990), Mangat and Singh (1990), Scheers (1992), Tracy and Mangat (1996), Singh et al. (2000) and Chaudhuri (2001) made further efforts to protect a respondent's privacy and increase response rates.

Common to these RR techniques is a sample drawn from the population by simple random sampling with or without replacement. Here randomized response is developed for a stratified random sampling. Stratified random sampling is generally obtained by dividing the population into non-overlapping groups called strata and selecting a simple random sample from each stratum. There are several reasons to apply randomized response to stratified random sampling. A randomized response using a stratified random sampling might give some clue to solve a limitation of randomized response which is the loss of individual characteristics of the respondents. By using the previous randomized response techniques, a group characteristic not individual data is obtained. A RR technique using a stratified random sampling gives the group characteristics related to each stratum estimator. For example, if strata are sex and age group, individual estimators for sex and age group answers can be obtained. The second reason to use stratified samples is that a researcher can be protected from the possibility of obtaining a really bad sample. Furthermore an administrative convenience reduces the cost of a stratified random sampling compared to a simple random sampling.

3.2. A Drawback of the Previous Stratified Randomized Response Model

Hong et al. (1994) suggested a stratified RR technique under the assumption that $n_i = n(N_i/N)$ where n_i and N_i are the sample size and the population size of stratum i, and n and N are the size of the whole sample and the size of the whole population. They applied the same randomization device that consists of a sensitive question (S) card with probability P and its negative question $(\bar{S})$ card with probability $1-P$ to every stratum. Under the proportional sampling assumption, it may be easy to derive the variance of the proposed estimator but may cause a high cost because of the difficulty in obtaining a proportional sample from some strata. To rectify this problem, a stratified randomized response technique using an optimal allocation is presented. It will be shown that a stratified randomized response technique using an optimal allocation is more efficient than a stratified randomized response technique using a proportional allocation.

3.3. Proposed Model

In the proposed model, the population is partitioned into strata, and a sample is selected by simple random sampling with replacement in each stratum. To get the full benefit from stratification, we assume that the number of units in each stratum is known. An individual respondent in the sample of stratum i is instructed to use the randomization device R_i which consists of a sensitive question (S) card with probability P_i and its negative question $(\bar{S})$ card with probability $1-P_i$. The respondent should answer the question by “Yes” or “No” without reporting which question card she or he has. This protects the respondent’s privacy. So a respondent belonging to the sample in different strata will use different randomization devices, each having different pre-assigned probabilities. Let n_i denote the number of units in the sample from stratum i and n denote the total number of units in samples from all strata so that $n=\sum_{i=1}^{k} n_i$.

Under the assumption that these “Yes” and “No” reports are made truthfully and $P_i(\neq 0.5)$ is set by the researcher, the probability of a “Yes” answer in a stratum i for this procedure is

$$Z_i = P_i\pi_{S_i} + (1-P_i)(1-\pi_{S_i}) \qquad \text{for } i=1,2,\cdots,k \tag{3.3.1}$$

where Z_i is the proportion of “Yes” answer in a stratum i, π_{S_i} is the proportion of respondents with the sensitive trait in a stratum i and P_i is the probability that a respondent in the sample in a stratum i has a sensitive question (S) card.

The maximum likelihood estimate of π_{S_i} is

$$\hat{\pi}_{S_i} = \frac{\hat{Z}_i - (1-P_i)}{2P_i - 1} \qquad \text{for } i=1,2,\cdots,k \tag{3.3.2}$$

where $\hat{Z}_i$ is the proportion of “Yes” answer in a sample in the stratum i and $\hat{\pi}_{S_i}$ is the proportion of respondents with the sensitive trait in a sample of the stratum i.

Since each $\hat{Z}_i$ is a binomial distribution $B(n_i, Z_i)$, the estimator $\hat{\pi}_{S_i}$ is unbiased for π_{S_i} with

$$Var(\hat{\pi}_{S_i}) = \frac{\pi_{S_i}(1-\pi_{S_i})}{n_i} + \frac{P_i(1-P_i)}{n_i(2P_i-1)^2}\,. \tag{3.3.3}$$

Since the selections in different strata are made independently, the estimators for individual strata can be added together to obtain an estimator for the whole population. The maximum likelihood estimate of π_S is easily shown to be

$$\hat{\pi}_S = \sum_{i=1}^{k} w_i \hat{\pi}_{S_i} = \sum_{i=1}^{k} w_i \left[\frac{\hat{Z}_i - (1-P_i)}{2P_i - 1} \right] \tag{3.3.4}$$

where we denote N to be the number of units in the whole population , N_i to be the total number of units in the stratum i and $w_i = (N_i/N)$ for $i = 1,2,\cdots,k$ so that $w = \sum_{i=1}^{k} w_i = 1$.

Theorem 3.3.1. The proposed estimator $\hat{\pi}_S$ is unbiased for population proportion π_S.

Proof. As each estimator $\hat{\pi}_{S_i}$ is unbiased for π_{S_i}, the expected value of $\hat{\pi}_S$ is

$$E(\hat{\pi}_S) = E\left(\sum_{i=1}^{k} w_i \hat{\pi}_{S_i} \right) = \sum_{i=1}^{k} w_i E(\hat{\pi}_{S_i}) = \sum_{i=1}^{k} w_i \pi_{S_i} = \pi_S .$$

The estimator $\hat{\pi}_S$ of π_S is unbiased.

Theorem 3.3.2. The variance of an estimator $\hat{\pi}_S$ is given by

$$Var(\hat{\pi}_S) = \sum_{i=1}^{k} \frac{w_i^2}{n_i} \left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i - 1)^2} \right]. \tag{3.3.5}$$

Proof. Since each unbiased estimator $\hat{\pi}_{S_i}$ has its own variance, the variance of $\hat{\pi}_S$ using (3.3.3) and corollary 1. in Sec. 5.9 of Cochran (1977) is

$$Var(\hat{\pi}_S) = Var\left(\sum_{i=1}^{k} w_i \hat{\pi}_{S_i} \right) = \sum_{i=1}^{k} w_i^2 Var(\hat{\pi}_{S_i}) = \sum_{i=1}^{k} \frac{w_i^2}{n_i} \left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i - 1)^2} \right]$$

which proves the theorem.

Information on π_{S_i} is usually unavailable. But if prior information on π_{S_i} is available from past experience then it helps to derive the following optimal allocation formula.

Theorem 3.3.3. The optimal allocation of n to $n_1, n_2, \cdots, n_{k-1}$ and n_k to derive the minimum variance of the $\hat{\pi}_S$ subject to $n = \sum_{i=1}^{k} n_i$ is approximately given by

$$\frac{n_i}{n} = \frac{w_i \left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i - 1)^2} \right]^{1/2}}{\sum_{i=1}^{k} w_i \left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i - 1)^2} \right]^{1/2}} . \tag{3.3.6}$$

Proof. For minimum variance for fixed total sample size in Sec. 5.9 of Cochran (1977),

$$n_i \propto N_i \left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i - 1)^2} \right]^{1/2} .$$

Thus

$$\frac{n_i}{n} \doteq \frac{N_i \left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i - 1)^2} \right]^{1/2}}{\sum_{i=1}^{k} N_i \left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i - 1)^2} \right]^{1/2}} = \frac{w_i \left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i - 1)^2} \right]^{1/2}}{\sum_{i=1}^{k} w_i \left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i - 1)^2} \right]^{1/2}} .$$

The proportion of the total sample size which is allocated to each sample is

$$\frac{n_i}{n} = \frac{w_i \left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i - 1)^2} \right]^{1/2}}{\sum_{i=1}^{k} w_i \left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i - 1)^2} \right]^{1/2}} .$$

Corollary 3.1. If we insert (3.3.6) into the following inequality:

$$\left[\sum_{i=1}^{k}\frac{w_i^2}{n_i}\left\{\pi_{S_i}(1-\pi_{S_i})+\frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}\right]\left(\sum_{i=1}^{k}n_i\right)\geq\left[\sum_{i=1}^{k}w_i\sqrt{\pi_{S_i}(1-\pi_{S_i})+\frac{P_i(1-P_i)}{(2P_i-1)^2}}\right]^2 \quad (3.3.7)$$

then we can easily show

$$\left[\sum_{i=1}^{k}\frac{w_i^2}{n_i}\left\{\pi_{S_i}(1-\pi_{S_i})+\frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}\right]\left(\sum_{i=1}^{k}n_i\right)=\left[\sum_{i=1}^{k}w_i\sqrt{\pi_{S_i}(1-\pi_{S_i})+\frac{P_i(1-P_i)}{(2P_i-1)^2}}\right]^2 \quad (3.3.8).$$

Proof. Suppose $k=3$. It means that $n=n_1+n_2+n_3$. Let

$$A=\sqrt{\pi_{S_1}(1-\pi_{S_1})+\frac{P_1(1-P_1)}{(2P_1-1)^2}},\ B=\sqrt{\pi_{S_2}(1-\pi_{S_2})+\frac{P_2(1-P_2)}{(2P_2-1)^2}}$$

$$\text{and}\quad C=\sqrt{\pi_{S_3}(1-\pi_{S_3})+\frac{P_3(1-P_3)}{(2P_3-1)^2}}.$$

By (3.3.6), we can derive the following ones:

$$\frac{n_1}{n_2}=\frac{w_i\sqrt{A}}{w_2\sqrt{B}}\quad\text{and}\quad\frac{n_3}{n_2}=\frac{w_3\sqrt{C}}{w_2\sqrt{B}}.$$

Thus

$$w_i\sqrt{A}=\frac{n_1}{n_2}w_2\sqrt{B}\quad\text{and}\quad w_3\sqrt{C}=\frac{n_3}{n_2}w_2\sqrt{B}.$$

We can use the above equations to show

$$\left[\sum_{i=1}^{3}\frac{w_i^2}{n_i}\left\{\pi_{S_i}(1-\pi_{S_i})+\frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}\right]\left(\sum_{i=1}^{k}n_i\right)=\left[\sum_{i=1}^{3}w_i\sqrt{\pi_{S_i}(1-\pi_{S_i})+\frac{P_i(1-P_i)}{(2P_i-1)^2}}\right]^2.$$

$$\left[\sum_{i=1}^{3} w_i \sqrt{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}}\right]^2 = \left[w_1\sqrt{A} + w_2\sqrt{B} + w_3\sqrt{C}\right]^2$$

$$= \left[\frac{n_1}{n_2} w_2\sqrt{B} + w_2\sqrt{B} + \frac{n_3}{n_2} w_2\sqrt{B}\right]^2 = w_2^2 B\left(\frac{n_1+n_2+n_3}{n_2}\right)^2 = w_2^2 B\left(\frac{n}{n_2}\right)^2$$

$$= \frac{n}{n_2^2}(w_2^2 B)n = \left(\frac{n_1+n_2+n_3}{n_2^2}\right)(w_2^2 B)n = \left(\frac{1}{n_1}\frac{n_1^2}{n_2^2} + \frac{1}{n_2} + \frac{1}{n_3}\frac{n_3^2}{n_2^2}\right)(w_2^2 B)n$$

$$= \left(\frac{1}{n_1}\frac{n_1^2}{n_2^2} w_2^2 B + \frac{1}{n_2} w_2^2 B + \frac{1}{n_3}\frac{n_3^2}{n_2^2} w_2^2 B\right)n = \left(\frac{1}{n_1} w_1^2 A + \frac{1}{n_2} w_2^2 B + \frac{1}{n_3}\frac{n_3^2}{n_2^2} w_3^2 C\right)n$$

$$= \left[\sum_{i=1}^{3} \frac{w_i^2}{n_i}\left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}\right]\left(\sum_{i=1}^{3} n_i\right).$$

By the mathematical induction, we can prove

$$\left[\sum_{i=1}^{k} \frac{w_i^2}{n_i}\left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}\right]\left(\sum_{i=1}^{k} n_i\right) = \left[\sum_{i=1}^{k} w_i \sqrt{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}}\right]^2$$

On using (3.3.8), we derive the minimal variance of an estimator $\hat{\pi}_S$ in the following theorem.

Theorem 3.3.4. The minimal variance of the estimator $\hat{\pi}_S$ is given by

$$Var(\hat{\pi}_S) = \frac{1}{n}\left[\sum_{i=1}^{k} w_i \left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}^{1/2}\right]^2. \tag{3.3.9}$$

Proof. From (3.3.8), we can derive the following equation:

$$\sum_{i=1}^{k} \frac{w_i^2}{n_i}\left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\} = \frac{1}{n}\left[\sum_{i=1}^{k} w_i \left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}^{1/2}\right]^2.$$

Inserting the right side of equation into (3.3.5),

$$Var(\hat{\pi}_S) = \sum_{i=1}^{k} \frac{w_i^2}{n_i}\left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} w_i \left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}^{1/2}\right]^2$$

which proves the Theorem.

Theorem 3.3.5. The unbiased estimator of the variance $Var(\hat{\pi}_S)$ is given by

$$v(\hat{\pi}_S) = \frac{1}{n-k}\left[\sum_{i=1}^{k} w_i \left\{\hat{\pi}_{S_i}(1-\hat{\pi}_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}^{1/2}\right]^2. \tag{3.3.10}$$

Proof. Substituting $n_i - 1$ for n_i in (3.3.5). Then

$$v(\hat{\pi}_S) = \sum_{i=1}^{k} \frac{w_i^2}{n_i - 1}\left[\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right]$$

and applying it to (3.3.6) and (3.3.8). Then

$$\left[\sum_{i=1}^{k} \frac{w_i^2}{n_i - 1}\left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}\right]\left(\sum_{i=1}^{k}(n_i - 1)\right) = \left[\sum_{i=1}^{k} w_i \sqrt{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}}\right]^2.$$

From the above equation,

$$\left[\sum_{i=1}^{k} \frac{w_i^2}{n_i - 1}\left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}\right] = \frac{1}{n-k}\left[\sum_{i=1}^{k} w_i \sqrt{\pi_{S_i}(1-\pi_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}}\right]^2.$$

Hence,

$$v(\hat{\pi}_S) = \frac{1}{n-k}\left[\sum_{i=1}^{k} w_i \left\{\hat{\pi}_{S_i}(1-\hat{\pi}_{S_i}) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}^{1/2}\right]^2.$$

3.4. Efficiency Comparison

3.4.1. Efficiency Comparison with the Hong et al. Model

Hong et al. (1994) derived the unbiased maximum likelihood estimator of π_S:

$$\hat{\pi}_H = \sum_{i=1}^{k} w_i \hat{\pi}_{S_i} = \sum_{i=1}^{k} w_i \left(\frac{\hat{Y}_i - (1-P)}{2P-1} \right) \tag{3.4.1.1}$$

where $\hat{Y}_i$ is the proportion of "Yes" answer in a sample in the stratum i. For its variance,

$$Var(\hat{\pi}_H) = \frac{1}{n} \sum_{i=1}^{k} \left[w_i \pi_{S_i}(1-\pi_{S_i}) \right] + \frac{P(1-P)}{n(2P-1)^2}$$

$$= \frac{1}{n} \left[\sum_{i=1}^{k} \left\{ w_i \pi_{S_i}(1-\pi_{S_i}) + \frac{P(1-P)}{(2P-1)^2} \right\} \right]. \tag{3.4.1.2}$$

under the assumption that $n_i = n(N_i/N)$.

From (3.3.9), we can get the variance of an estimator of our proposed stratified randomized response technique.

Suppose $P_i = P$ for all i. Then (3.3.9) becomes

$$Var(\hat{\pi}_S) = \frac{1}{n} \left[\sum_{i=1}^{k} w_i \left\{ \pi_{S_i}(1-\pi_{S_i}) + \frac{P(1-P)}{(2P-1)^2} \right\}^{1/2} \right]^2. \tag{3.4.1.3}$$

We can do a mathematical comparison as follows:

We denote $L_i = \sqrt{\pi_{S_i}(1-\pi_{S_i}) + \frac{P(1-P)}{(2P-1)^2}}$. We can do the following:

$$Var(\hat{\pi}_H) - Var(\hat{\pi}_S) = \frac{1}{n}\left(\sum_{i=1}^{k} w_i L_i^2 \right) - \frac{1}{n}\left(\sum_{i=1}^{k} w_i L_i \right)^2 = \frac{1}{n}\left[\left(\sum_{i=1}^{k} w_i L_i^2 \right) - \left(\sum_{i=1}^{k} w_i L_i \right)^2 \right]$$

$$Var(\hat{\pi}_H) - Var(\hat{\pi}_S) = \frac{1}{n}\left[\left(\sum_{i=1}^{k} w_i L_i^2\right) - 2\left(\sum_{i=1}^{k} w_i L_i\right)^2 + \left(\sum_{i=1}^{k} w_i L_i\right)^2\right]$$

$$= \frac{1}{n}\left[\left(\sum_{i=1}^{k} w_i L_i^2\right) - 2\left(\sum_{i=1}^{k} w_i L_i\right)\left(\sum_{i=1}^{k} w_i L_i\right) + \left(\sum_{i=1}^{k} w_i L_i\right)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{k} w_i\left\{L_i^2 - 2L_i\left(\sum_{i=1}^{k} w_i L_i\right) + \left(\sum_{i=1}^{k} w_i L_i\right)^2\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{k} w_i\left\{L_i - \left(\sum_{i=1}^{k} w_i L_i\right)\right\}^2$$

which is always positive. Therefore, $Var(\hat{\pi}_H) - Var(\hat{\pi}_S) > 0$.

The relative efficiency of two variances is

$$\text{R.E.} = \frac{Var(\hat{\pi}_H)}{Var(\hat{\pi}_S)} > 1.$$

Since the value of the R.E. is more than one, our proposed stratified RR technique is more efficient than the Hong et al. (1994) stratified RR technique when $P_i = P$ for all i. In a case that $P_i \neq P$ for all i, it is difficult to derive the mathematical condition of the relative efficiency comparison from (3.3.9) and (3.4.1.2). We resort to the empirical study on the percent relative efficiency (RE). Suppose that there are two strata in a population and $P_2 > P_1$ from (3.3.9). Then the percent relative efficiency is

$$\text{Percent RE} = \frac{Var(\hat{\pi}_H)}{Var(\hat{\pi}_S)} \times 100. \tag{3.4.1.4}$$

If the value of the percent RE is more than 100 then our proposed model is more efficient than the Hong et al. (1994) model. But if the percent RE is less then 100, then the Hong et al. model is more efficient than the proposed model. Since a sample size n is cancelled out in the percent RE, we do not have to change the sample size in the percent RE. Suppose that we can get prior information on π_{S_1}, π_{S_2}, w_1, w_2, π_S. Under the condition $P_2 > P_1$, Table 3.1 shows that the values of the percent relative efficiency are more than 100 for all parameter values tabled. We obtain the values of the percent relative efficiency from changing π_{S_1}, π_{S_2}, w_1, w_2, $n = 1000$ and P_2. Since the Warner model is symmetric in terms of P, the values of the percent relative efficiency are also symmetric in terms of P. We just showed the cases from $P = 0.6$ to $P = 0.9$ by 0.1 increments. We dealt with the empirical study of the percent relative efficiency of $Var(\hat{\pi}_H)/Var(\hat{\pi}_S)$ in the case of two strata. In section 3.4.3, we will think about more than two strata cases in terms of efficiency. We will verify that we will have the same result in more than two strata as that for two strata in the population. From the two cases presented in this paper, we may conclude that our stratified randomized response technique is more efficient than the stratified randomized response technique presented by Hong et al. (1994).

TABLE 3.1.

The Percent Relative Efficiency of $Var(\hat{\pi}_H)/Var(\hat{\pi}_S)$ When $n=1000$.

					$P=P_1$							
					0.6		0.7		0.8		0.9	
					P_2		P_2		P_2		P_2	
π_{S_1}	π_{S_2}	w_1	w_2	π_S	0.7	0.8	0.8	0.9	0.9	0.95	0.93	0.95
0.08	0.13	0.7	0.3	0.095	140.2	160.2	127.2	147	123.5	134.3	107.5	112.5
		0.6	0.4	0.1	159.1	192.8	138.8	170.1	133.1	149.5	110.1	117
		0.4	0.6	0.11	210.4	296.6	166.9	235.8	155.6	188.2	115.3	126.4
		0.3	0.7	0.115	245.9	383.1	184.3	283.8	168.8	213.1	118	131.5
0.18	0.23	0.7	0.3	0.195	139.4	158.3	125.4	142.6	120	128.3	105.6	109.2
		0.6	0.4	0.2	157.8	189.5	136	162.9	127.9	140.4	107.6	112.5
		0.4	0.6	0.21	207.4	287.3	161.6	219	146.1	169.9	111.6	119.5
		0.3	0.7	0.215	241.5	367.4	177.1	258.3	156.6	188.1	113.6	123.1
0.28	0.33	0.7	0.3	0.295	138.8	157.1	124.2	140.1	118.1	125.3	104.8	107.8
		0.6	0.4	0.3	156.9	187.4	134.3	158.8	125.1	135.8	106.5	110.6
		0.4	0.6	0.31	205.4	281.5	158.3	209.7	141.3	161.2	109.9	116.4
		0.3	0.7	0.315	238.6	357.7	172.8	244.7	150.5	176.5	111.6	119.5
0.38	0.43	0.7	0.3	0.395	138.5	156.4	123.6	138.7	117.1	123.8	104.4	107.2
		0.6	0.4	0.4	156.4	186.2	133.4	156.8	123.8	133.7	106	109.7
		0.4	0.6	0.41	204.3	278.5	156.7	205.1	139	157.2	109.1	115.1
		0.3	0.7	0.415	237	352.7	170.6	238.1	147.6	171.3	110.7	117.9
0.48	0.53	0.7	0.3	0.495	138.4	156.3	123.5	138.4	116.9	123.5	104.3	107
		0.6	0.4	0.5	156.3	186	133.2	156.3	123.5	133.2	105.8	109.5
		0.4	0.6	0.51	204.1	277.9	156.3	204.2	138.5	156.4	108.9	114.8
		0.3	0.7	0.515	236.7	351.7	170.2	236.9	147	170.3	110.6	117.6
0.58	0.63	0.7	0.3	0.595	138.6	156.6	123.8	139.1	117.3	124.2	104.5	107.3
		0.6	0.4	0.6	156.5	186.6	133.6	157.4	124.1	134.2	106.1	109.9
		0.4	0.6	0.61	204.7	279.7	157.2	206.8	139.7	158.5	109.3	115.5
		0.3	0.7	0.615	237.7	354.8	171.5	240.7	148.6	173.2	111	118.5
0.68	0.73	0.7	0.3	0.695	139	157.5	124.5	140.8	118.5	126.1	105	108.1
		0.6	0.4	0.7	157.2	188.2	134.8	160.2	125.9	137.1	106.7	111.1
		0.4	0.6	0.71	206.2	284.1	159.5	213.3	143	164.3	110.4	117.4
		0.3	0.7	0.715	239.9	362.2	174.6	250.5	152.9	181.1	112.3	120.8

					$P=P_1$							
					0.6		0.7		0.8		0.9	
					P_2		P_2		P_2		P_2	
π_{S_1}	π_{S_2}	w_1	w_2	π_S	0.7	0.8	0.8	0.9	0.9	0.95	0.93	0.95
0.78	0.83	0.7	0.3	0.795	139.6	159	125.9	144	120.8	129.9	106	109.9
		0.6	0.4	0.8	158.3	190.8	136.9	165.4	129.4	143.1	108.1	113.5
		0.4	0.6	0.81	208.7	291.5	163.6	225.8	149.5	176.4	112.7	121.6
		0.3	0.7	0.815	243.5	374.8	180.1	269.3	161.4	197.8	115.1	126
0.88	0.93	0.7	0.3	0.895	140.6	161.3	128.1	149.5	125.3	137.8	108.4	114.1
		0.6	0.4	0.9	159.8	194.8	140.2	174.6	136	155.6	111.5	119.7
		0.4	0.6	0.91	212.3	303	170.3	248.8	162.5	203.9	118.2	132.3
		0.3	0.7	0.915	248.8	394.6	189.3	305.2	178.9	237.4	121.9	139.6

3.4.2. Efficiency Comparison with Variations of the Warner Model

We will do an efficiency comparison of our stratified randomized response technique and two-stage randomized response technique that was presented by Mangat and Singh (1990) by a way of variance comparison.

Theorem 3.4.1. Suppose that there are two strata in the population and $P=P_1=P_2\neq 0.5$. The proposed estimator $\hat{\pi}_S$ will be more efficient than the Mangat and Singh (1990) estimator $\hat{\pi}_{ms}$ under the following condition:

$$(\pi_{S_1}-\pi_{S_2})^2+\left[\left\{\pi_{S_1}(1-\pi_{S_1})+\frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}-\left\{\pi_{S_2}(1-\pi_{S_2})+\frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2 \quad (3.4.2.1)$$

$$>\left[\left\{\frac{M(1-P)}{(2P-1)(2P-1+2M(1-P))}\right\}^2-\frac{M(1-P)}{(1-2P)(2P-1+2M(1-P))^2}\right]\left[w_1(1-w_1)\right]^{-1}$$

where $\pi_{S_1}\neq\pi_{S_2}$.

Proof. Assume $n = n_1 + n_2$, $P = P_1 = P_2 \neq 0.5$ and $\hat{\pi}_S = w_1\hat{\pi}_{S_1} + w_2\hat{\pi}_{S_2}$.

On using (2.1.14) and (3.3.9), we check an efficiency of $\hat{\pi}_S$ with respect to $\hat{\pi}_{ms}$.

$$Var(\hat{\pi}_{ms}) - Var(\hat{\pi}_S) = \frac{\pi_S(1-\pi_S)}{n}$$

$$+\frac{(1-M)(1-P)\{1-(1-M)(1-P)\}}{n\{2P-1+2M(1-P)\}^2} - \frac{1}{n}\left[\sum_{i=1}^{2} w_i \left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2.$$

Inserting $\pi_S = w_1\pi_{S_1} + w_2\pi_{S_2}$ into $Var(\hat{\pi}_{ms})$, then we can derive the following equation:

$$Var(\hat{\pi}_{ms}) - Var(\hat{\pi}_S) = \frac{(w_1\pi_{S_1} + w_2\pi_{S_2}) - (w_1^2\pi_{S_1}^2 + w_2^2\pi_{S_2}^2) - 2w_1w_2\pi_{S_1}\pi_{S_2}}{n}$$

$$+\frac{(1-M)(1-P)\{1-(1-M)(1-P)\}}{n\{2P-1+2M(1-P)\}^2} - \frac{1}{n}\left[\sum_{i=1}^{2} w_i \left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2.$$

$$Var(\hat{\pi}_{ms}) - Var(\hat{\pi}_S) = \frac{w_1\pi_{S_1} + w_2\pi_{S_2} - 2w_1w_2\pi_{S_1}\pi_{S_2}}{n} + \frac{(1-M)(1-P)\{1-(1-M)(1-P)\}}{n\{2P-1+2M(1-P)\}^2}$$

$$-\left(\frac{w_1^2\pi_{S_1} + w_2^2\pi_{S_2}}{n}\right) - \frac{(w_1^2 + w_2^2)P(1-P)}{n(2P-1)^2}$$

$$-\frac{2w_1(1-w_1)}{n}\left[\pi_{S_1}(1-\pi_{S_1}) + \frac{P(1-P)}{(2P-1)^2}\right]^{1/2}\left[\pi_{S_2}(1-\pi_{S_2}) + \frac{P(1-P)}{(2P-1)^2}\right]^{1/2}.$$

Since

$$\frac{w_1\pi_{S_1} + w_2\pi_{S_2} - 2w_1w_2\pi_{S_1}\pi_{S_2}}{n} - \left(\frac{w_1^2\pi_{S_1} + w_2^2\pi_{S_2}}{n}\right) = \frac{w_1(1-w_1)(\pi_{S_1} + \pi_{S_2} - 2\pi_{S_1}\pi_{S_2})}{n},$$

We can derive the following equation:

$$Var(\hat{\pi}_{ms}) - Var(\hat{\pi}_S) = \frac{w_1(1-w_1)(\pi_{S_1} - \pi_{S_2})^2}{n}$$

$$+\frac{w_1(1-w_1)}{n}\left[\left\{\pi_{S_1}(1-\pi_{S_1}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2} - \left\{\pi_{S_2}(1-\pi_{S_2}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2$$

$$+\frac{(1-M)(1-P)\{1-(1-M)(1-P))\}}{n\{2P-1+2M(1-P)\}^2} - \frac{P(1-P)}{n(2P-1)^2} \quad .$$

After some algebra, we can derive the following:

$$Var(\hat{\pi}_{ms}) - Var(\hat{\pi}_S) = \frac{w_1(1-w_1)(\pi_{S_1} - \pi_{S_2})^2}{n} - \frac{M[M(1-P)^2 - 1 + 3P - 2P^2]}{n(2P-1)^2[2P-1+2M(1-P)]^2}$$

$$+\frac{w_1(1-w_1)}{n}\left[\left\{\pi_{S_1}(1-\pi_{S_1}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2} - \left\{\pi_{S_2}(1-\pi_{S_2}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2 .$$

Since $\dfrac{M[M(1-P)^2 - 1 + 3P - 2P^2]}{n(2P-1)^2[2P-1+2M(1-P)]^2}$

$$= \frac{1}{n}\left[\frac{M(1-P)}{(2P-1)(2P-1+2M(1-P))}\right]^2 - \frac{M(1-P)}{n(1-2P)[2P-1+2M(1-P)]^2} \;,$$

$$Var(\hat{\pi}_{ms}) - Var(\hat{\pi}_S) = \frac{w_1(1-w_1)(\pi_{S_1} - \pi_{S_2})^2}{n}$$

$$+\frac{w_1(1-w_1)}{n}\left[\left\{\pi_{S_1}(1-\pi_{S_1}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2} - \left\{\pi_{S_2}(1-\pi_{S_2}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2$$

$$-\frac{1}{n}\left[\frac{M(1-P)}{(2P-1)(2P-1+2M(1-P))}\right]^2 + \frac{M(1-P)}{n(1-2P)[2P-1+2M(1-P)]^2} .$$

If $Var(\hat{\pi}_{ms}) - Var(\hat{\pi}_S) > 0$ then the proposed estimator $\hat{\pi}_S$ will be more efficient than that of Mangat and Singh (1990).

In a case of $\pi_{S_1} \neq \pi_{S_2}$,

$$(\pi_{S_1} - \pi_{S_2})^2 + \left[\left\{ \pi_{S_1}(1-\pi_{S_1}) + \frac{P(1-P)}{(2P-1)^2} \right\}^{1/2} - \left\{ \pi_{S_2}(1-\pi_{S_2}) + \frac{P(1-P)}{(2P-1)^2} \right\}^{1/2} \right]^2$$

$$> \left[\left\{ \frac{M(1-P)}{(2P-1)(2P-1+2M(1-P))} \right\}^2 - \frac{M(1-P)}{(1-2P)(2P-1+2M(1-P))^2} \right] [w_1(1-w_1)]^{-1} .$$

If prior information on π_{S_1}, π_{S_2}, w_1, w_2 and n can be roughly obtained and M and $P = P_1 = P_2 \neq 0.5$ are chosen by the researcher, then we can easily check the relative efficiency of $Var(\hat{\pi}_{ms}) / Var(\hat{\pi}_S)$. Suppose we have prior information on π_{S_1}, π_{S_2}, w_1, w_2 and n. Then we set four different P's and three different M's to verify the relative efficiency of $\hat{\pi}_S$ with respect to $\hat{\pi}_{ms}$ in Table 3.2. Under the condition (3.4.2.1), Table 3.2 shows that the proposed estimator $\hat{\pi}_S$ is more efficient than the Mangat and Singh (1990) estimator $\hat{\pi}_{ms}$. Warner (1965) mentioned that a P close to 1 (or close to 0) is adequate to insure cooperation from respondents but a value of P close to 0.5 conveys less information from each interview. Thus four different P's and three different M's we used in Table 3.2 are adequate to insure cooperation. From the Table 3.2, we can make several observations. The first observation is that every value in Table 3.2 is much bigger than one, indicating that the relative efficiency of the proposed method is considerably higher than that of Mangat and Singh (1990). The second observation is that the value of relative efficiency increases as P and M increase, except for $P = 0.35$ with $M = 0.3$, $P = 0.4$ with $M = 0.3$, and $P = 0.4$ with $M = 0.2$ in every case. The third observation is that when $M = 0.3$ and $P = 0.3$, the value of relative efficiency is unusually high in every case. An additional observation is that there is little reduction in relative efficiency as π_S increases.

In the empirical investigation, we do not change sample size n in the Table 3.2 because n is cancelled out in the ratio of $Var(\hat{\pi}_{ms})/Var(\hat{\pi}_S)$. Through these results, we have demonstrated that the proposed estimator $\hat{\pi}_S$ be more efficient than that of Mangat and Singh (1990) under (3.4.2.1) in a case of two strata in the population. When $M = 0.1$, the Figure 3.1 shows that the relative efficiency of $\hat{\pi}_S$ with respect to $\hat{\pi}_{ms}$ increases as P increases, but there is little reduction of the relative efficiency as π_S increases.

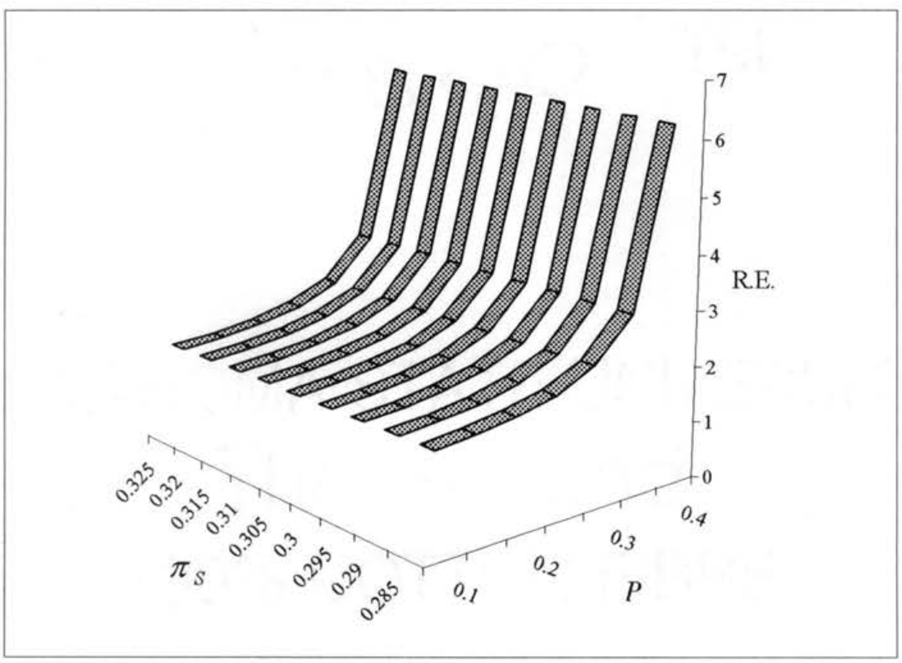

Figure3.1. The Relative Efficiency of $Var(\hat{\pi}_{ms})/Var(\hat{\pi}_S)$ When $M = 0.1$

TABLE 3.2.

The Relative Efficiency of $Var(\hat{\pi}_{ms})/Var(\hat{\pi}_{s})$ When $n=1000$ and $P=P_1=P_2\neq 0.5$.

						P						
π_{S_1}	π_{S_2}	w_1	w_2	π_S	M	0.1	0.15	0.2	0.25	0.3	0.35	0.4
0.28	0.33	0.9	0.1	0.285	0.1	1.7554	1.8195	1.9215	2.0918	2.4089	3.1502	6.2893
					0.2	3.6179	4.06	4.8498	6.506	11.421	57.19	25.18
					0.3	10.612	14.834	26.721	104.82	412.23	11.283	1.5667
0.28	0.33	0.8	0.2	0.29	0.1	1.7518	1.8165	1.919	2.0897	2.4072	3.1488	6.2877
					0.2	3.6038	4.0477	4.8385	6.495	11.408	57.15	25.172
					0.3	10.559	14.777	26.645	104.61	411.7	11.276	1.5666
0.28	0.33	0.7	0.3	0.295	0.1	1.748	1.8134	1.9165	2.0877	2.4055	3.1473	6.2861
					0.2	3.5897	4.0354	4.8273	6.484	11.395	57.11	25.165
					0.3	10.506	14.72	26.569	104.4	411.17	11.269	1.5664
0.28	0.33	0.6	0.4	0.3	0.1	1.7441	1.8102	1.9139	2.0855	2.4037	3.1458	6.2844
					0.2	3.5756	4.023	4.816	6.473	11.381	57.07	25.157
					0.3	10.453	14.663	26.493	104.19	410.65	11.261	1.5663
0.28	0.33	0.5	0.5	0.305	0.1	1.7401	1.8069	1.9112	2.0834	2.402	3.1443	6.2828
					0.2	3.5614	4.0105	4.8047	6.4619	11.368	57.03	25.149
					0.3	10.401	14.607	26.417	103.98	410.12	11.254	1.5661
0.28	0.33	0.4	0.6	0.31	0.1	1.7359	1.8035	1.9085	2.0811	2.4002	3.1427	6.2811
					0.2	3.5472	3.9981	4.7934	6.4508	11.355	56.99	25.142
					0.3	10.349	14.551	26.341	103.77	409.6	11.247	1.5659
0.28	0.33	0.3	0.7	0.315	0.1	1.7316	1.8001	1.9057	2.0789	2.3983	3.1412	6.2794
					0.2	3.533	3.9856	4.7821	6.4397	11.342	56.95	25.134
					0.3	10.297	14.495	26.266	103.57	409.08	11.239	1.5657
0.28	0.33	0.2	0.8	0.32	0.1	1.7272	1.7965	1.9028	2.0766	2.3964	3.1396	6.2778
					0.2	3.5187	3.973	4.7707	6.4286	11.328	56.91	25.127
					0.3	10.246	14.439	26.191	103.36	408.56	11.232	1.5655
0.28	0.33	0.1	0.9	0.325	0.1	1.7227	1.7928	1.8998	2.0742	2.3946	3.138	6.2761
					0.2	3.5044	3.9605	4.7593	6.4174	11.315	56.871	25.119
					0.3	10.195	14.384	26.117	103.15	408.04	11.225	1.5653

Prior information on π_{S_1}, π_{S_2}, w_1, w_2, $\pi_S = w_1\pi_{S_1} + w_2\pi_{S_2}$, n, M and P satisfy the condition (3.4.2.1).

If we set $M = 0$ in the two-stage RR model presented by the Mangat and Singh (1990), then the Mangat and Singh (1990) method reduces to the Warner (1965) model.

Theorem 3.4.2. Suppose there are two strata in the population and $P = P_1 = P_2 \neq 0.5$. The proposed estimator $\hat{\pi}_S$ is more efficient than the Warner (1965) estimator $\hat{\pi}_W$.

Proof. Suppose $M = 0$ in the Mangat and Singh (1990) model. We can show that $Var(\hat{\pi}_{ms}) = Var(\hat{\pi}_W)$.

From the condition (3.4.2.1) when $M = 0$,

$$(\pi_{S_1} - \pi_{S_2})^2 + \left[\left\{ \pi_{S_1}(1-\pi_{S_1}) + \frac{P(1-P)}{(2P-1)^2} \right\}^{1/2} - \left\{ \pi_{S_2}(1-\pi_{S_2}) + \frac{P(1-P)}{(2P-1)^2} \right\}^{1/2} \right]^2 > 0$$

where $\pi_{S_1} \neq \pi_{S_2}$.

$$Var(\hat{\pi}_W) - Var(\hat{\pi}_S) = \frac{w_1(1-w_1)(\pi_{S_1} - \pi_{S_2})^2}{n}$$

$$+ \frac{w_1(1-w_1)}{n} \left[\left\{ \pi_{S_1}(1-\pi_{S_1}) + \frac{P(1-P)}{(2P-1)^2} \right\}^{1/2} - \left\{ \pi_{S_2}(1-\pi_{S_2}) + \frac{P(1-P)}{(2P-1)^2} \right\}^{1/2} \right]^2 > 0.$$

The proposed estimator $\hat{\pi}_S$ is always more efficient than the Warner estimator $\hat{\pi}_w$ when $\pi_{S_1} \neq \pi_{S_2}$.

We showed that the proposed estimator is more efficient than that of the Warner (1965). Mangat (1994) also showed that his estimator is more efficient than that of Warner (1965) if $\pi_S > 1 - \{P/(2P-1)\}^2$ which always holds for $P > 1/3$. The following theorem is to compare two different estimators under his condition with respect to efficiency.

Theorem 3.4.3. Suppose that $\pi_S > 1-\{P/(2P-1)\}^2$ and assume that there are two strata in the population and $P = P_1 = P_2 \neq 0.5$. The proposed estimator $\hat{\pi}_S$ will be more efficient than the Mangat (1994) estimator $\hat{\pi}_m$ under the following condition:

$$(\pi_{S_1} - \pi_{S_2})^2 + \left[\left\{\pi_{S_1}(1-\pi_{S_1}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2} - \left\{\pi_{S_2}(1-\pi_{S_2}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2$$

$$> \frac{(1-P)}{w_1(1-w_1)P}\left[\left(\frac{P}{2P-1}\right)^2 - \{1-(w_1\pi_{S_1} + w_2\pi_{S_2})\}\right] \qquad (3.4.2.2)$$

where $\pi_{S_1} \neq \pi_{S_2}$.

Proof. Assume $\pi_S > 1-\{P/(2P-1)\}^2$, $n = n_1 + n_2$ and $P = P_1 = P_2 \neq 0.5$. By using (2.1.16) and (3.3.9),

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_S) = \frac{\pi_S(1-\pi_S)}{n} + \frac{(1-\pi_S)(1-P)}{nP} - \frac{1}{n}\left[\sum_{i=1}^{2} w_i\left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2$$

.

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_S) = \frac{\pi_S(1-\pi_S)}{n} + \frac{P(1-P)}{n(2P-1)^2}$$

$$+ \frac{(1-\pi_S)(1-P)}{nP} - \frac{P(1-P)}{n(2P-1)^2} - \frac{1}{n}\left[\sum_{i=1}^{2} w_i\left\{\pi_{S_i}(1-\pi_{S_i}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2.$$

We can derive

$$\frac{(1-\pi_S)(1-P)}{nP} - \frac{P(1-P)}{n(2P-1)^2} = \frac{(1-P)}{nP}\left[(1-\pi_S) - \left(\frac{P}{2P-1}\right)^2\right].$$

Since $\pi_S = w_1\pi_{S_1} + w_2\pi_{S_2}$, we derive the following:

$$Var(\hat{\pi}_m)-Var(\hat{\pi}_S)=\frac{(w_1\pi_{S_1}+w_2\pi_{S_2})[1-(w_1\pi_{S_1}+w_2\pi_{S_2})]}{n}+\frac{P(1-P)}{n(2P-1)^2}$$

$$-\frac{(1-P)}{nP}\left[\left(\frac{P}{2P-1}\right)^2-\{1-(w_1\pi_{S_1}+w_2\pi_{S_2})\}\right]-\frac{1}{n}\left[\sum_{i=1}^{2}w_i\left\{\pi_{S_i}(1-\pi_{S_i})+\frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2$$

$$=\frac{(w_1\pi_{S_1}+w_2\pi_{S_2})-(w_1^2\pi_{S_1}^2+w_2^2\pi_{S_2}^2)-2w_1w_2\pi_{S_1}\pi_{S_2}}{n}+\frac{P(1-P)}{n(2P-1)^2}$$

$$-\frac{1}{n}\left[\sum_{i=1}^{2}w_i\left\{\pi_{S_i}(1-\pi_{S_i})+\frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2-\frac{(1-P)}{nP}\left[\left(\frac{P}{2P-1}\right)^2-\{1-(w_1\pi_{S_1}+w_2\pi_{S_2})\}\right].$$

$$Var(\hat{\pi}_m)-Var(\hat{\pi}_S)=\frac{w_1(1-w_1)}{n}\left[\pi_{S_1}+\pi_{S_2}-2\pi_{S_1}\pi_{S_2}+\frac{2P(1-P)}{(2P-1)^2}\right]$$

$$-\frac{2w_1(1-w_1)}{n}\left[\pi_{S_1}(1-\pi_{S_1})+\frac{P(1-P)}{(2P-1)^2}\right]^{1/2}\left[\pi_{S_2}(1-\pi_{S_2})+\frac{P(1-P)}{(2P-1)^2}\right]^{1/2}$$

$$-\frac{(1-P)}{nP}\left[\left(\frac{P}{2P-1}\right)^2-\{1-(w_1\pi_{S_1}+w_2\pi_{S_2})\}\right].$$

$$Var(\hat{\pi}_m)-Var(\hat{\pi}_S)=\frac{w_1(1-w_1)}{n}\left[\pi_{S_1}+\pi_{S_2}-2\pi_{S_1}\pi_{S_2}+\frac{2P(1-P)}{(2P-1)^2}\right]$$

$$-\frac{w_1(1-w_1)}{n}\left[\left\{\pi_{S_1}(1-\pi_{S_1})+\frac{P(1-P)}{(2P-1)^2}\right\}+\left\{\pi_{S_2}(1-\pi_{S_2})+\frac{P(1-P)}{(2P-1)^2}\right\}\right]$$

$$+\frac{w_1(1-w_1)}{n}\left[\left\{\pi_{S_1}(1-\pi_{S_1})+\frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}-\left\{\pi_{S_2}(1-\pi_{S_2})+\frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2$$

$$-\frac{(1-P)}{nP}\left[\left(\frac{P}{2P-1}\right)^2-\{1-(w_1\pi_{S_1}+w_2\pi_{S_2})\}\right].$$

Therefore

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_S) = \frac{w_1(1-w_1)(\pi_{S_1} - \pi_{S_2})^2}{n}$$

$$+ \frac{w_1(1-w_1)}{n}\left[\left[\left\{\pi_{S_1}(1-\pi_{S_1}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2} - \left\{\pi_{S_2}(1-\pi_{S_2}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2\right]$$

$$- \frac{(1-P)}{nP}\left[\left(\frac{P}{2P-1}\right)^2 - \{1-(w_1\pi_{S_1} + w_2\pi_{S_2})\}\right].$$

By the assumption $\pi_S > 1 - \{P/(2P-1)\}^2$,

$$\left[\left(\frac{P}{2P-1}\right)^2 - \{1-(w_1\pi_{S_1} + w_2\pi_{S_2})\}\right] > 0.$$

To show that the proposed estimator $\hat{\pi}_S$ is more efficient than the Mangat (1994) estimator, $Var(\hat{\pi}_m) - Var(\hat{\pi}_S)$ should be positive. Using this fact, we can derive the following condition:

$$(\pi_{S_1} - \pi_{S_2})^2 + \left[\left[\left\{\pi_{S_1}(1-\pi_{S_1}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2} - \left\{\pi_{S_2}(1-\pi_{S_2}) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2\right]$$

$$> \frac{(1-P)}{w_1(1-w_1)P}\left[\left(\frac{P}{2P-1}\right)^2 - \{1-(w_1\pi_{S_1} + w_2\pi_{S_2})\}\right]$$

where $\pi_{S_1} \neq \pi_{S_2}$.

We showed that our proposed estimator will be more efficient than the previous three estimators under the conditions (3.4.2.1) and (3.4.2.2) by a variance comparison in a case of two strata in the population.

3.4.3. Cost and Efficiency of Stratification

We need think about more than two strata cases in terms of efficiency. Cochran (1977) showed that the variance for the mean of a stratified random sample decreases as the number of strata increases. So we want to show that the variance of an estimator in our RR model decreases as the number of strata increases.

Suppose that k strata of equal size are created such that $w_i = 1/k$. Inserting $w_i = 1/k$ into equation (3.3.9), then

$$Var(\hat{\pi}_S) = \frac{1}{nk^2}\left[\sum_{i=1}^{k}\sqrt{\pi_{S_i}(1-\pi_{S_i}) + \left(P_i(1-P_i)/(2P_i-1)^2\right)}\right]^2. \qquad (3.4.3.1)$$

Let $f(k) = \frac{1}{k^2}\left[\sum_{i=1}^{k}\sqrt{\pi_{S_i}(1-\pi_{S_i}) + \left(P_i(1-P_i)/(2P_i-1)^2\right)}\right]^2$ where k is a positive integer. We want to show that $f(k) - f(k+1) \geq 0$.

For $L(\pi_{S_i}, P_i) = \sqrt{\pi_{S_i}(1-\pi_{S_i}) + \left(P_i(1-P_i)/(2P_i-1)^2\right)}$,

$$f(k) - f(k+1) = \frac{1}{k^2}\left[\sum_{i=1}^{k} L(\hat{\pi}_{S_i}, P_i)\right]^2 - \frac{1}{(k+1)^2}\left[\sum_{i=1}^{k+1} L(\hat{\pi}_{S_i}, P_i)\right]^2$$

$$= \left[\frac{1}{k}\left(\sum_{i=1}^{k} L(\hat{\pi}_{S_i}, P_i)\right) + \frac{1}{k+1}\left(\sum_{i=1}^{k+1} L(\hat{\pi}_{S_i}, P_i)\right)\right]\left[\frac{1}{k}\left(\sum_{i=1}^{k} L(\hat{\pi}_{S_i}, P_i)\right) - \frac{1}{k+1}\left(\sum_{i=1}^{k+1} L(\hat{\pi}_{S_i}, P_i)\right)\right].$$

As the number of strata increases, it may be possible to divide a heterogeneous population into subpopulations, each of which is more homogeneous. So we may get

$$\left[\frac{1}{k}\left(\sum_{i=1}^{k} L(\hat{\pi}_{S_i}, P_i)\right) - \frac{1}{k+1}\left(\sum_{i=1}^{k+1} L(\hat{\pi}_{S_i}, P_i)\right)\right] \geq 0.$$

By this assumption, $f(k)$ is a monotone decreasing function of k. Thus the variance of an estimator decreases as the number of strata increases. Therefore, the variance of our proposed estimator will be smaller as the number of strata increases.

Next question we have is how much the value of a variance will decrease as the number of strata increases. Kish (1965) answered our question by quoting the following model $R^2/I^2 + (1-R^2)$ where R^2 is the portion of the variance affected by the stratification and I is the number of strata. By this model, he wrote "the variance approaches to $(1-R^2)$ after the creation of a moderate number of strata". Thus, little reduction in variance will be expected beyond an adequate number of strata in the population. We can tell that the cost of the survey affected by an increase of the number of strata is the limitation of a stratified random sampling method. Thus our proposed model has the same limitation. We recommend that when a researcher wants to increase the number of strata in a population, she or he should consider carefully whether the decrease of the variance (increase in precision) is worth the extra cost involved in increasing the number of strata. However, since a researcher may get a gain in precision in the estimates of the sensitive trait proportion in the population and can compare the target groups in which she or he is interested, a stratified RR model is an advantageous model compared to the RR model using simple random sampling.

3.5. Less Than Completely Truthful Reporting

We denote T_r to be the weighted probability $T_r = \sum_{i=1}^{k} w_i T_{r_i}$ where T_{r_i} is the probability that a respondent with the sensitive trait will report truthfully in a sample stratum i. We assume that the respondents with the non-sensitive trait will report truthfully. The probability of a "Yes" answer in a stratum i for this procedure is given by

$$Z_i' = P_i \pi_{S_i} T_r + (1-P_i)\pi_{S_i}(1-T_r) + (1-P_i)(1-\pi_{S_i}) \qquad \text{where} \quad i = 1,2,\cdots,k \qquad (3.5.1)$$

A biased estimator $\hat{\pi}'_S$ of π_S in the population has the following bias and variance:

$$Bias(\hat{\pi}'_S) = E(\hat{\pi}'_S - \hat{\pi}_S) = \sum_{i=1}^{k} w_i E(\hat{\pi}'_{S_i} - \hat{\pi}_{S_i}) = \sum_{i=1}^{k} w_i \pi_{S_i}(T_r - 1). \tag{3.5.2}$$

$$Var(\hat{\pi}'_S) = \frac{1}{n}\left[\sum_{i=1}^{k} w_i \left\{\pi_{S_i} T_r (1-\pi_{s_i} T_r) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}^{1/2}\right]^2. \tag{3.5.3}$$

The mean square error of $\hat{\pi}'_S$ is given by

$$MSE(\hat{\pi}'_S) = \frac{1}{n}\left[\sum_{i=1}^{k} w_i \left\{\pi_{S_i} T_r (1-\pi_{s_i} T_r) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}^{1/2}\right]^2 + \left\{\sum_{i=1}^{k} w_i \pi_{S_i}(T_r - 1)\right\}^2. \tag{3.5.4}$$

The following theorem compares the efficiency of the proposed estimator $\hat{\pi}'_S$ and the Mangat and Singh (1990) estimator $\hat{\pi}'_{ms}$ in a situation of less than completely truthful reporting.

Theorem 3.5.1. Suppose there are two strata in the population and $P = P_1 = P_2 \neq 0.5$. The proposed estimator $\hat{\pi}'_S$ will be more efficient than Mangat and Singh (1990) estimator $\hat{\pi}'_{ms}$ if

$$(\pi_{S_1} T_r - \pi_{S_2} T_r)^2 + \left[\left\{\pi_{S_1} T_r (1-\pi_{s_{1i}} T_r) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2} - \left\{\pi_{S_2} T_r (1-\pi_{s_2} T_r) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2$$

$$> \frac{2P(1-P)}{(2P-1)^2} - \frac{(1-M)(1-P)\{1-(1-M)(1-P)\}}{w_1(1-w_1)\{2P-1+2M(1-P)\}^2}$$

$$-\pi_S M(T-T_r)[1+\pi_S(n-1)\{M(T-T_r)+4MT_r(1-P)+2T_r(2P-1)\}$$

$$-2(1-M)(1-P)-2\pi_S n\{2M(1-P)+2P-1\}][w_1(1-w_1)\{2P-1+2M(1-P)\}^2]^{-1}.$$

(3.5.5)

Proof. From (2.1.15), we can get the mean square error of $\hat{\pi}'_{ms}$ from Mangat and Singh (1990):

$$MSE(\hat{\pi}'_{ms}) = \frac{\pi_S T_r (1-\pi_S T_r)}{n} + \frac{(1-M)(1-P)[1-(1-M)(1-P)]}{n[2P-1+2M(1-P)]^2} + [\pi_S (T_r - 1)]^2$$

$$+\pi_S M(T-T_r)[1+\pi_S(n-1)\{M(T-T_r)+4MT_r(1-P)+2T_r(2P-1)\}$$

$$-2(1-M)(1-P)-2\pi_S n\{2M(1-P)+2P-1\}][n\{2P-1+2M(1-P)\}^2]^{-1}$$

where T and T_r are the probabilities that a respondent with the sensitive trait will report truthfully at the first stage and second stage. From (3.5.4), the mean square error of $\hat{\pi}'_S$ is

$$MSE(\hat{\pi}'_S) = \frac{1}{n}\left[\sum_{i=1}^{2} w_i \left\{\pi_{S_i} T_r (1-\pi_{S_i} T_r) + \frac{P_i(1-P_i)}{(2P_i-1)^2}\right\}^{1/2}\right]^2 + \left\{\sum_{i=1}^{2} w_i \pi_{S_i} (T_r - 1)\right\}^2 .$$

For $\pi_S = w_1\pi_{S_1} + w_2\pi_{S_2}$, the difference of two mean square errors of $\hat{\pi}'_S$ and $\hat{\pi}'_{ms}$ is

$$MSE(\hat{\pi}'_{ms}) - MSE(\hat{\pi}'_S) = \frac{w_1(1-w_1)}{n}\left[(\pi_{S_1}T_r - \pi_{S_2}T_r)^2 + \left\{\left(\pi_{S_1}T_r(1-\pi_{S_{1i}}T_r) + \frac{P(1-P)}{(2P-1)^2}\right)^{1/2}\right.\right.$$

$$\left.\left. - \left(\pi_{S_2}T_r(1-\pi_{S_2}T_r) + \frac{P(1-P)}{(2P-1)^2}\right)^{1/2}\right\}^2\right] - \frac{2w_1(1-w_1)P(1-P)}{n(2P-1)^2} + \frac{(1-M)(1-P)\{1-(1-M)(1-P)\}}{n\{2P-1+2M(1-P)\}^2}$$

$$+\pi_S M(T-T_r)[1+\pi_S(n-1)\{M(T-T_r)+4MT_r(1-P)+2T_r(2P-1)\}$$

$$-2(1-M)(1-P)-2\pi_S n\{2M(1-P)+2P-1\}][n\{2P-1+2M(1-P)\}^2]^{-1}.$$

The proposed estimator $\hat{\pi}'_S$ will be more efficient than the Mangat and Singh (1990) estimator $\hat{\pi}'_{ms}$ if $MSE(\hat{\pi}'_S) < MSE(\hat{\pi}'_{ms})$.

$$(\pi_{S_1}T_r - \pi_{S_2}T_r)^2 + \left[\left\{\pi_{S_1}T_r(1-\pi_{s_{1i}}T_r) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2} - \left\{\pi_{S_2}T_r(1-\pi_{s_2}T_r) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2$$

$$> \frac{2P(1-P)}{(2P-1)^2} - \frac{(1-M)(1-P)\{1-(1-M)(1-P)\}}{w_1(1-w_1)\{2P-1+2M(1-P)\}^2}$$

$$-\pi_S M(T-T_r)[1+\pi_S(n-1)\{M(T-T_r)+4MT_r(1-P)+2T_r(2P-1)\}$$

$$-2(1-M)(1-P)-2\pi_S n\{2M(1-P)+2P-1\}][w_1(1-w_1)\{2P-1+2M(1-P)\}^2]^{-1}$$

which proves (3.5.5). We derive the following one from $MSE(\hat{\pi}'_{ms}) - MSE(\hat{\pi}'_S) > 0$.

If a researcher could obtain prior information on $\pi_{S_1}, \pi_{S_2}, w_1, w_2, n$ and M, then a researcher can check a relative efficiency of $MSE(\hat{\pi}'_{ms}) / MSE(\hat{\pi}'_S)$ with prior information of T and T_r.

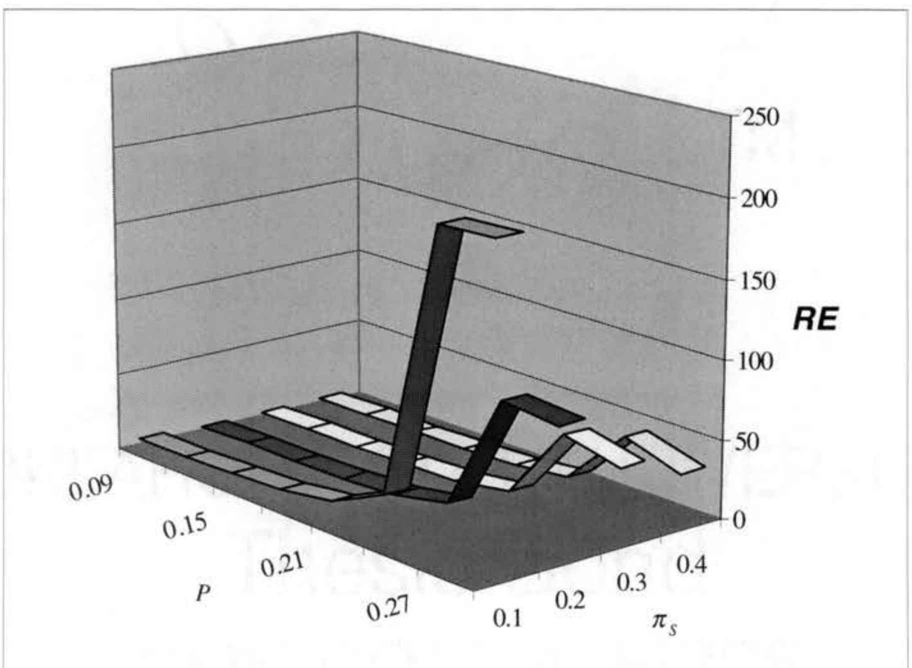

Figure 3.2. The Relative Efficiency of $MSE(\hat{\pi}'_{ms}) / MSE(\hat{\pi}'_S)$

When $T = 0.8, T_r = 0.7$, $M = 0.3$ and $n = 2000$.

Under prior information on π_{S_1}, π_{S_2}, w_1, w_2, M and differing levels of n, P, T and T_r satisfying (3.5.1), Table 3.3 shows that the proposed estimator $\hat{\pi}'_S$ is more efficient than the Mangat and Singh (1990) estimator $\hat{\pi}'_{ms}$ in the case with two strata in terms of the relative efficiency, $MSE(\hat{\pi}'_{ms}) / MSE(\hat{\pi}'_S)$.

When $T = 0.8$, $T_r = 0.7$, $M = 0.3$ and $n = 2000$, Figure 3.2 shows that the value of relative efficiency is decreasing as π_S increases, but the value of the relative efficiency is increasing as P increases. Table 3.3 and Fig. 3.2 show that our proposed estimator is more efficient than that of Mangat and Singh (1990) under condition (3.5.1). In a case of $M = 0$ in (2.1.15), $MSE(\hat{\pi}'_{ms})$ reduces to $MSE(\hat{\pi}'_w)$. So an efficiency comparison of the proposed estimator and that of Warner (1965) in a situation of less than completely truthful reporting is given by the following theorem.

Theorem 3.5.2. The proposed estimator $\hat{\pi}'_S$ is more efficient than the Warner (1965) estimator $\hat{\pi}'_W$ in the case of two strata in the population and $P = P_1 = P_2 \neq 0.5$.

Proof. The proof is similar to that of Theorem 4.2.

Suppose $M = 0$ in the equation (2.1.15) of the Mangat and Singh (1990) model. It is shown that $MSE(\hat{\pi}'_{ms}) = MSE(\hat{\pi}'_W)$. From the condition (3.5.1) when $M = 0$,

$$(\pi_{S_1}T_r - \pi_{S_2}T_r)^2 + \left[\left\{\pi_{S_1}T_r(1-\pi_{S_{1i}}T_r) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2} - \left\{\pi_{S_2}T_r(1-\pi_{S_2}T_r) + \frac{P(1-P)}{(2P-1)^2}\right\}^{1/2}\right]^2$$
$$> \frac{2P(1-P)}{(2P-1)^2} - \frac{P(1-P)}{w_1(1-w_1)(2P-1)^2}$$

where $\pi_{S_1} \neq \pi_{S_2}$.

The difference of two mean square errors of $\hat{\pi}'_s$ and $\hat{\pi}'_w$ is

$$MSE(\hat{\pi}'_w) - MSE(\hat{\pi}'_s) = \frac{w_1(1-w_1)}{n}\left[(\pi_{s_1}T_r - \pi_{s_2}T_r)^2 + \left\{\left(\pi_{s_1}T_r(1-\pi_{s_{1i}}T_r) + \frac{P(1-P)}{(2P-1)^2}\right)^{1/2}\right.\right.$$

$$\left.\left. - \left(\pi_{s_2}T_r(1-\pi_{s_2}T_r) + \frac{P(1-P)}{(2P-1)^2}\right)^{1/2}\right\}^2\right] + \frac{[1-2w_1(1-w_1)]P(1-P)}{n(2P-1)^2} > 0 .$$

It means that the proposed estimator is always more efficient than that of Warner in a situation of less than completely truthful reporting when there are two strata in the population.

Remark. The mean square error of our proposed estimator can be compared with that of Mangat (1994):

$$MSE(\hat{\pi}'_m) = \frac{\pi_S T_r(1-\pi_S T_r)}{nP^2} + \frac{(1-\pi_S)(1-P)[1-(1-\pi_S)(1-P) - 2\pi_S T_r]}{nP^2} + \left[\frac{\pi_S(T_r - 1)}{P}\right]^2$$

for an efficiency comparison in a case of less than completely truthful reporting.

We showed that our proposed estimator is more efficient than the Warner and the Mangat and Singh estimators under the condition reference in the case with two strata in the population when the respondents are not completely truthful reporting in their answers. Using Cochran's result for a stratified random sampling, we can insist that our proposed estimator should be more efficient than the Warner and the Mangat and Singh (1990) estimators in a case of less than completely truthful reporting.

3.6. Discussion

This paper presented a new stratified randomized response model. We showed that our model is more efficient than the Hong et al. (1994) stratified randomized response model. In both situations of the completely truthful reporting and less than completely truthful reporting, we showed that the proposed randomized response model is more efficient than the Warner (1965), the Mangat and Singh (1990) and the Mangat (1994) randomized response model with the condition presented. With prior information satisfying the conditions (3.4.2.1) and (3.5.1), we showed the relative efficiency of the proposed estimator with respect to the Mangat and Singh (1990) estimator. Tables 3.2, 3.3 and Figures 3.1, 3.2 show that the relative efficiency is very high under the condition reference. Furthermore, the proposed method is more useful than the previous methods in that a stratified randomized response method helps to solve the limitation of randomized response that is the loss of individual characteristics of the respondents. Therefore, the proposed method has several advantages compared to the previous randomized response methods.

CHAPTER IV

A MIXED RANDOMIZED RESPONSE MODEL

4.1. Introduction

For socially undesirable questions, direct measurement of valid information on human populations is difficult because of non-sampling errors, that is, refusal to respond and untruthful reporting. The randomized response (RR) survey technique that Warner (1965) proposed for the first time is designed to encourage cooperation and truthful replies to questions involving socially undesirable activities. A common objective of all randomized response variants of the Warner model is the protection of privacy while improving accuracy by reduction in response bias. Researchers such as Horvitz et al. (1967), Greenberg et al. (1969), Moors (1971), Lanke (1975, 1976), Anderson (1976), Leysieffer and Warner (1976), Greenberg et al. (1977), Flinger et al. (1977), Chaudhuri and Mukerjee (1988), Kuk (1990), Ljungqvist (1993), Mangat et al. (1993), Nayak (1994), Mangat et al. (1997), Singh et al. (2000) made an effort to protect a respondent's privacy and increase response rates by deriving the optimal design of RR model. The researchers compared RR designs based on statistical measure of efficiency and respondents' protection.

4.2. A Privacy Problem of the Moors' Model

Mangat et al. (1997) and Singh et al. (2000) pointed out the privacy problem of the Moors model. They assumed that a respondent belongs to the sensitive trait group but does not belong to the innocuous trait group. Suppose that the respondent is independently chosen in two samples drawn from the population using simple random sampling with replacement. If the respondent chosen in the first sample must answer Question A, then his or her answer should be "Yes". By this assumption, the respondent is also chosen in the second sample. So if the respondent in the second sample must answer Question B, then his or her answer should be "No". Thus, the respondent common to both samples answered "Yes" when he or she had Question A and "No" when having Question B. Hence, interviewer can determine that the respondent belongs to the sensitive trait group. The privacy of the respondent is not protected in the Moors' model. As an alternative model of the Moors model, Mangat et al. (1997) proposed a random group method. The method can protect respondents' privacy but there is an efficiency problem. Mangat et al. (1997) mentioned that the variance yielded by the random group method is greater than that for the Moors model. Singh et al. (2000) proposed two different models as alternatives for the Moors model. But these two models have a common weak point. This weak point is that although their models using simple random sampling without replacement might be more efficient than the Moors model while keeping the confidentiality of a respondent, those models lead to a high cost survey since those alternative models need larger sample sizes than the Moors model using simple random sampling with replacement. Thus these drawbacks with the previous alternative models for the Moors model motivate the authors to propose another alternative model that will rectify the problems presented in the above models.

4.3. Proposed model

4.3.1. A Background of Deriving a New RR Model

Fox and Tracy (1986) described the choice of a nonsensitive question like this; "The respondent reporting sensitive information is provided very little protection from a small π_I, largely defeating the purpose of using a randomized response. Whenever π_I approaches 0, the conditional probability of having the sensitive attribute given a "Yes" answer, $P(A \mid Yes)$, is uncomfortably high." Lanke (1975) demonstrated that under the condition $P(A \mid Yes) <$ constant, the standard deviation of the unrelated question RR model when π_I is known is a decreasing function of π_I. Hence, under this condition, the RR design of $\pi_I = 1$ yields the minimum value of the variance for the proposed estimator and helps to minimize the risk of suspicion when the respondent possessing a sensitive trait responds "Yes". Furthermore, the respondent possessing a innocuous trait feels more comfortable to answer "Yes" in the design of $\pi_I = 1$. Despite these advantages of choosing $\pi_I = 1$, Greenberg et al. (1977) disagreed with the idea of Lanke (1975) because the expected overall benefit of randomized response technique will be zero in the case of the randomized response design of $\pi_I = 1$. But this does not affect to our mixed RR model. The reason will be discussed in detail in Section 4.3.3.

4.3.2. A Mixed Randomized Response Model

In this proposed model, a single sample with size n is selected by simple random sampling with replacement from the population. Each respondent from the sample is instructed to answer a direct question about "I am a member of the innocuous trait group". If a respondent answers "Yes", then she or he is instructed to go to a randomization device R_1 consisting of the two statements. One statement is "I am a member of the sensitive trait group", and the other one is "I am a member of the innocuous trait group" with preassigned probability of selections of P_1 and $1 - P_1$ respectively. If a respondent answers "No", then the respondent is instructed to use a randomization device R_2 consisting of the two statements. One statement is "I am a member of the sensitive trait group", and the other one is "I am not a member of the sensitive trait group" with preassigned probabilities P and $1-P$ respectively. Thus the Warner model requires that the innocuous question be the same at both steps in the process. To protect respondents' privacy, the respondents should not disclose the question they answered from either randomization R_1 or R_2 to the interviewer. The proportion of "Yes" answer from the respondents using randomization device R_1 is

$$Y_1 = P_1\pi_S + (1-P_1)\pi_I. \tag{4.3.2.1}$$

Since the respondent using a randomization device R_1 already responded "Yes" from the initial direct innocuous question, π_I is equal to one. Therefore, (4.3.2.1) becomes $Y_1 = P_1\pi_S + (1-P_1)$. The estimate of π_S, in terms of sample proportions of "Yes" responses, $\hat{Y}_1$, becomes

$$\hat{\pi}_{UY} = \frac{\hat{Y}_1 - (1-P_1)}{P_1}. \tag{4.3.2.2}$$

For its variance,

$$Var(\hat{\pi}_{UY}) = \frac{Y_1(1-Y_1)}{n_1P_1^2} = \frac{[P_1\pi_S + (1-P_1)][1-P_1\pi_S-(1-P_1)]}{n_1P_1^2}$$

$$= \frac{P_1(1-\pi_S)[P_1\pi_S+(1-P_1)]}{n_1P_1^2} = \frac{(1-\pi_S)[P_1\pi_S+(1-P_1)]}{n_1P_1}. \tag{4.3.2.3}$$

where n_1 is the number of people responding "Yes" when respondents in a sample n were asked the direct innocuous question. An unbiased estimate of $Var(\hat{\pi}_{UY})$ is

$$v(\hat{\pi}_{UY}) = \frac{\hat{Y}_1(1-\hat{Y}_1)}{(n_1-1)P_1^2} = \frac{(1-\hat{\pi}_S)[P_1\hat{\pi}_S+(1-P_1)]}{(n_1-1)P_1}. \tag{4.3.2.4}$$

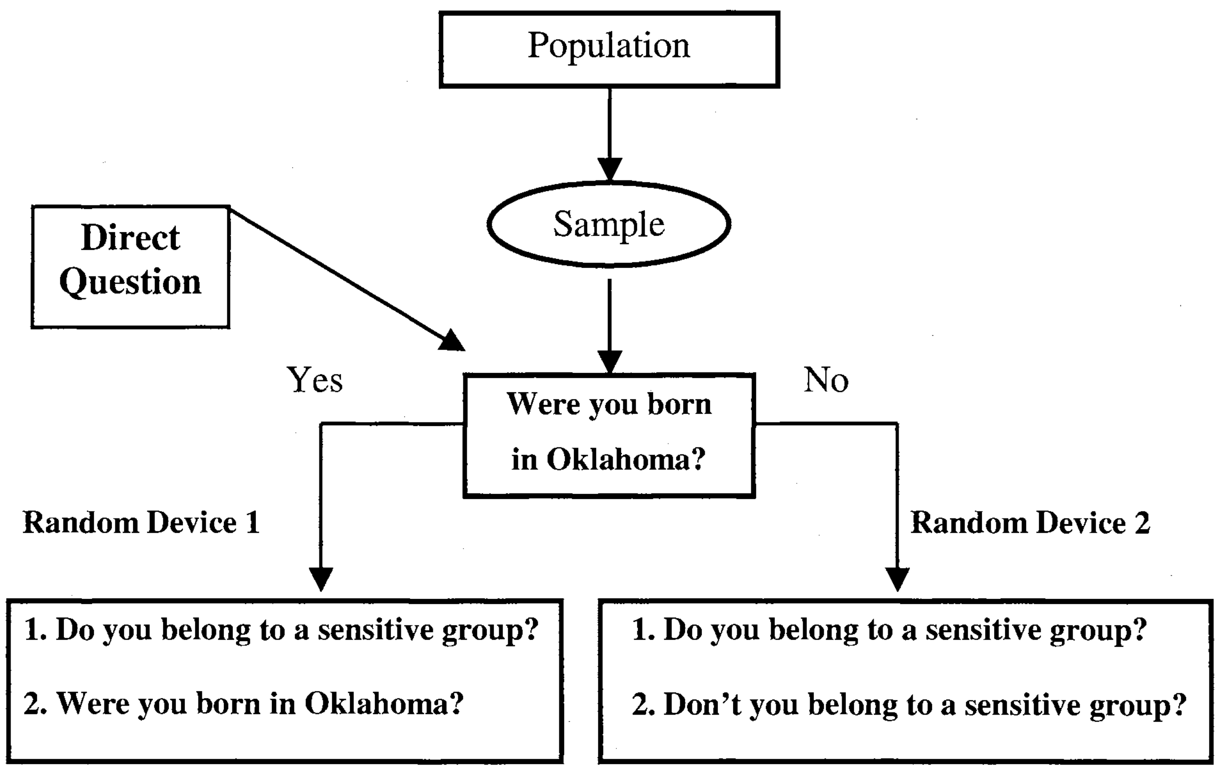

Figure 4.1. A Mixed Randomized Response Model

The proportion of "Yes" answer from the respondents using a randomization device R_2 is

$$X = P\pi_S + (1-P)(1-\pi_S) = (2P-1)\pi_S + 1 - P. \tag{4.3.2.5}$$

The estimator of π_S, in terms of sample proportions of "Yes" responses, $\hat{X}$, becomes

$$\hat{\pi}_W = \frac{\hat{X} - (1-P)}{2P-1} \tag{4.3.2.6}$$

with

$$Var(\hat{\pi}_W) = \frac{X(1-X)}{n_2 P^2} = \frac{\pi_S(1-\pi_S)}{n_2} + \frac{P(1-P)}{n_2(2P-1)^2}. \tag{4.3.2.7}$$

where n_2 is the number of people responding "No" when respondents in a sample n had the direct question. An unbiased estimate of $Var(\hat{\pi}_W)$ is

$$v(\hat{\pi}_W) = \frac{\hat{X}(1-\hat{X})}{(n_2-1)P^2} = \frac{\hat{\pi}_W(1-\hat{\pi}_W)}{n_2-1} + \frac{P(1-P)}{(n_2-1)(2P-1)^2}. \tag{4.3.2.8}$$

Then the estimator of π_S, in terms of sample proportions of "Yes" responses, $\hat{Y}_1$ and $\hat{X}$, as

$$\hat{\pi}_m = \frac{n_1}{n}\hat{\pi}_{UY} + \frac{n_2}{n}\hat{\pi}_W \qquad \text{for } 0 < \frac{n_1}{n} \text{ and } \frac{n_2}{n} < 1 \tag{4.3.2.9}$$

with

$$Var(\hat{\pi}_m) = Var\left(\frac{n_1}{n}\hat{\pi}_{UY} + \frac{n_2}{n}\hat{\pi}_W\right) = \left(\frac{n_1}{n}\right)^2 Var(\hat{\pi}_{UY}) + \left(\frac{n_2}{n}\right)^2 Var(\hat{\pi}_W)$$

$$= \left(\frac{n_1}{n}\right)^2 \left[\frac{(1-\pi_S)\{P_1\pi_S + (1-P_1)\}}{n_1 P_1}\right] + \left(\frac{n_2}{n}\right)^2 \left[\frac{\pi_S(1-\pi_S)}{n_2} + \frac{P(1-P)}{n_2(2P-1)^2}\right]. \tag{4.3.2.10}$$

Since the previous researchers showed that the unrelated question RR model is generally more efficient than the Warner model, we allocate more respondents to the unrelated question RR model in a mixed RR model than to the Warner model in a mixed RR model. Then we can make the variance of the estimator in our mixed RR model smaller.

Theorem 4.3.2.1. The proposed estimator $\hat{\pi}_m$ is unbiased for population proportion π_S.

Proof. As both $\hat{\pi}_{UY}$ and $\hat{\pi}_W$ are unbiased estimators, the expected value of $\hat{\pi}_m$ is

$$E(\hat{\pi}_m) = E\left(\frac{n_1}{n}\hat{\pi}_{UY} + \frac{n_2}{n}\hat{\pi}_W\right) = \frac{n_1}{n}E(\hat{\pi}_{UY}) + \frac{n_2}{n}E(\hat{\pi}_W)$$

$$= \frac{n_1}{n}\pi_{UY} + \frac{n_2}{n}\pi_W = \pi_S.$$

An estimator $\hat{\pi}_m$ of π_S is unbiased.

Theorem 4.3.2.2. When the Warner model and unrelated question RR model are equally protective, the variance of $\hat{\pi}_W$ can be expressed for every P_1 and $\pi_I = 1$ as:

$$Var(\hat{\pi}_W) = \frac{\pi_S(1-\pi_S)}{n_2} + \frac{1-P_1}{n_2 P_1^2}. \qquad (4.3.2.11)$$

Proof. Lanke (1976) derived a unique value of P as

$$P = \frac{1}{2} + \frac{P_1}{2P_1 + 4(1-P_1)\pi_I} \qquad \text{for every } P_1 \text{ and every } \pi_I$$

such that the Warner model and unrelated question RR model are equally confidential to respondents : $\quad P_W(A\,|\,Yes) = P_{UK}(A\,|\,Yes) \quad$ for every π_S.

Since there is a $\pi_I = 1$ in the mixed randomized response model,

$$P = \frac{1}{2} + \frac{P_1}{2P_1 + 4(1-P_1)} = \frac{1}{2-P_1}$$

Inserting $P=(2-P_1)^{-1}$ into (4.3.2.7). Then we can derive the following one:

$$Var(\hat{\pi}_W)=\frac{\pi_S(1-\pi_S)}{n_2}+\frac{P(1-P)}{n_2(2P-1)^2}=\frac{\pi_S(1-\pi_S)}{n_2}+\frac{1-P_1}{n_2P_1^2}$$

which proves the Theorem.

Theorem 4.3.2.3. The variance of an estimator $\hat{\pi}_m$ is given by

$$Var(\hat{\pi}_m)=\frac{\pi_S(1-\pi_S)}{n}+\frac{(1-P_1)[\lambda P_1(1-\pi_S)+(1-\lambda)]}{nP_1^2} \qquad (4.3.2.12)$$

where $n=n_1+n_2$ and $\lambda=n_1/n$.

Proof. Suppose $n=n_1+n_2$. Using (4.3.2.10) and (4.3.2.11) equations,

we can derive $Var(\hat{\pi}_m)$:

$$Var(\hat{\pi}_m)=Var\left(\frac{n_1}{n}\hat{\pi}_{UY}+\frac{n_2}{n}\hat{\pi}_W\right)=\left(\frac{n_1}{n}\right)^2Var(\hat{\pi}_{UY})+\left(\frac{n_2}{n}\right)^2Var(\hat{\pi}_W)$$

$$=\left(\frac{n_1}{n}\right)^2\left[\frac{(1-\pi_S)\{P_1\pi_S+(1-P_1)\}}{n_1P_1}\right]+\left(\frac{n_2}{n}\right)^2\left[\frac{\pi_S(1-\pi_S)}{n_2}+\frac{1-P_1}{n_2P_1^2}\right]$$

$$=\frac{1}{(nP_1)^2}\left[n_1P_1(1-\pi_S)\{P_1\pi_S+(1-P_1)\}+n_2P_1^2\pi_S(1-\pi_S)+n_2(1-P_1)\right]$$

$$=\frac{1}{(nP_1)^2}\left[n_1P_1(1-\pi_S)\{P_1\pi_S+(1-P_1)\}+n_2P_1^2\pi_S(1-\pi_S)+n_2(1-P_1)\right]$$

$$=\frac{1}{(nP_1)^2}\left[(n_1+n_2)P_1^2\pi_S(1-\pi_S)+(1-P_1)\{n_1P_1(1-\pi_S)+n_2\}\right]$$

$$=\frac{\pi_S(1-\pi_S)}{n}+\frac{(1-P_1)[n_1P_1(1-\pi_S)+n_2]}{n^2P_1^2}$$

$$=\frac{\pi_S(1-\pi_S)}{n}+\frac{(1-P_1)[\lambda P_1(1-\pi_S)+(1-\lambda)]}{nP_1^2} \qquad \text{where } \lambda=n_1/n.$$

Theorem 4.3.2.4. The unbiased variance of an estimator $\hat{\pi}_S$ is given by

$$v(\hat{\pi}_m) = \frac{\hat{\pi}_m(1-\hat{\pi}_m)}{n^2}\left[\frac{n_1^2}{n_1-1}+\frac{n_2^2}{n_2-1}\right]+\frac{(1-P_1)}{(nP_1)^2}\left[\frac{n_1^2 P_1(1-\hat{\pi}_m)}{n_1-1}+\frac{n_2^2}{n_2-1}\right]. \quad (4.3.2.13)$$

Proof. Using (4.3.2.4), (4.3.2.8) and (4.3.2.11) equations, we can derive the following one: For $n = n_1 + n_2$,

$$v(\hat{\pi}_m) = v\left(\frac{n_1}{n}\hat{\pi}_{UY} + \frac{n_2}{n}\hat{\pi}_W\right) = \left(\frac{n_1}{n}\right)^2 v(\hat{\pi}_{UY}) + \left(\frac{n_2}{n}\right)^2 v(\hat{\pi}_W)$$

$$= \left(\frac{n_1}{n}\right)^2\left[\frac{(1-\hat{\pi}_m)\{P_1\hat{\pi}_m + (1-P_1)\}}{(n_1-1)P_1}\right] + \left(\frac{n_2}{n}\right)^2\left[\frac{\hat{\pi}_m(1-\hat{\pi}_m)}{n_2-1}+\frac{1-P_1}{(n_2-1)P_1^2}\right]$$

$$= \frac{1}{(nP_1)^2}\left[\frac{n_1^2 P_1(1-\hat{\pi}_m)\{P_1\hat{\pi}_m+(1-P_1)\}}{n_1-1}+\frac{n_2^2\{P_1^2\hat{\pi}_m(1-\hat{\pi}_m)+(1-P_1)\}}{n_2-1}\right]$$

$$= \frac{1}{(nP_1)^2}\left[P_1^2\hat{\pi}_m(1-\hat{\pi}_m)\left\{\frac{n_1^2}{n_1-1}+\frac{n_2^2}{n_2-1}\right\}+\frac{n_1^2 P_1(1-P_1)(1-\hat{\pi}_m)}{n_1-1}+\frac{n_2^2(1-P_1)}{n_2-1}\right]$$

$$= \frac{\hat{\pi}_m(1-\hat{\pi}_m)}{n^2}\left[\frac{n_1^2}{n_1-1}+\frac{n_2^2}{n_2-1}\right]+\frac{(1-P_1)}{(nP_1)^2}\left[\frac{n_1^2 P_1(1-\hat{\pi}_m)}{n_1-1}+\frac{n_2^2}{n_2-1}\right]$$

which proves the Theorem.

4.3.3. A Validation of a Mixed RR Model.

We can determine the estimate of π_I from a direct question before performing the randomization devices R_1 and R_2 by asking a direct question about an innocuous trait. If the researcher avoids selecting the innocuous trait direct question so that all respondents answer "Yes" to it, then the criticism by Greenberg et al. (1977) in terms of $\pi_I = 1$ does not apply to our mixed RR model. From Greenberg et al. (1977) we obtain the expected overall benefit (EOB_1) from the unrelated question RR model when π_I is known. The expected overall benefit is

$$EOB_1 = \pi_S(1 - P_1)(1 - \pi_I). \tag{4.3.3.1}$$

We will derive the expected overall benefit of the Warner model from the respondent hazard concept of Greenberg et al. (1977). They defined the hazard for a respondent from a sensitive group S as the probability that the respondent in S is perceived as belonging to S,

$$H_S = P(Yes \mid S)P(S \mid Yes) + P(No \mid S)P(S \mid No).$$

Similarly, they defined the hazard for a respondent from a nonsensitive group $\bar{S}$ is the probability that the respondent in $\bar{S}$ is perceived as belonging to S,

$$H_{\bar{S}} = P(Yes \mid \bar{S})P(S \mid Yes) + P(No \mid \bar{S})P(S \mid No).$$

From the Warner model, we can derive the following:

$$P(Yes \mid S) = P(No \mid \bar{S}) = P \ , \ P(No \mid S) = P(Yes \mid \bar{S}) = 1 - P \ ,$$

$$P(S \mid Yes) = \frac{P\pi_S}{P\pi_S + (1-P)(1-\pi_S)}, \text{ and } P(S \mid No) = \frac{(1-P)\pi_S}{(1-P)\pi_S + P(1-\pi_S)}. \tag{4.3.3.2}$$

Therefore,

$$\begin{aligned} H_S &= P(Yes \mid S)P(S \mid Yes) + P(No \mid S)P(S \mid No) \\ &= P\frac{P\pi_S}{P\pi_S + (1-P)(1-\pi_S)} + (1-P)\frac{(1-P)\pi_S}{(1-P)\pi_S + P(1-\pi_S)} \\ &= \frac{P^2\pi_S}{P\pi_S + (1-P)(1-\pi_S)} + \frac{(1-P)^2\pi_S}{(1-P)\pi_S + P(1-\pi_S)}. \end{aligned} \tag{4.3.3.3}$$

$$H_{\bar{S}} = P(Yes \mid \bar{S})P(S \mid Yes) + P(No \mid \bar{S})P(S \mid No)$$

$$= (1-P)\frac{P\pi_S}{P\pi_S + (1-P)(1-\pi_S)} + P\frac{(1-P)\pi_S}{(1-P)\pi_S + P(1-\pi_S)}$$

$$= \frac{P(1-P)\pi_S}{P\pi_S + (1-P)(1-\pi_S)} + \frac{P(1-P)\pi_S}{(1-P)\pi_S + P(1-\pi_S)}. \tag{4.3.3.4}$$

They explained the limited hazard. It is likely to be closer to the actual concern felt by a respondent as the probability that a respondent in a sensitive group S answer "Yes" and is perceived as belonging to S, $LH_S = P(Yes \mid S)P(S \mid Yes)$, and the probability that a respondent in a sensitive group $\bar{S}$ answer "Yes" and is perceived as belonging to S, $LH_{\bar{S}} = P(Yes \mid \bar{S})P(S \mid Yes)$. Hence,

$$LH_S = \frac{P^2\pi_S}{P\pi_S + (1-P)(1-\pi_S)} \text{ and } LH_{\bar{S}} = \frac{P(1-P)\pi_S}{P\pi_S + (1-P)(1-\pi_S)}. \tag{4.3.3.5}$$

The expected overall benefit for the Warner model is

$$EOB_2 = \pi_S(1 - LH_S) + (1-\pi_S)(-LH_{\bar{S}})$$

$$= \pi_S\left[1 - \frac{P^2\pi_S}{P\pi_S + (1-P)(1-\pi_S)}\right] - (1-\pi_S)\left[\frac{P(1-P)\pi_S}{P\pi_S + (1-P)(1-\pi_S)}\right]$$

$$= \pi_S(1-P). \tag{4.3.3.6}$$

The expected overall benefit for a mixed RR model is

$$EOB = \frac{n_1}{n}EOB_1 + \frac{n_2}{n}EOB_2 = \frac{n_1}{n}\pi_S(1-P_1)(1-\pi_I) + \frac{n_2}{n}\pi_S(1-P). \tag{4.3.3.7}$$

Since there is a $\pi_I = 1$ in the unrelated question RR model part from our mixed RR model,

$$EOB = \frac{n_1}{n}EOB_1 + \frac{n_2}{n}EOB_2 = \frac{n_2}{n}\pi_S(1-P). \tag{4.3.3.8}$$

Since n_2/n is an estimate of $1-\pi_I$, $E(n_2/n)=1-\pi_I$. If $P_1=P$, the expected overall benefit for the proposed mixed RR model is close to the expected overall benefit for unrelated question RR model when π_I is known. Thus, we can conclude that the expected overall benefit of a $\pi_I=1$ design in our mixed RR model will not be zero. So a $\pi_I=1$ design in our mixed RR model may not be criticized by arguments such as these presented in Greenberg et al. (1977).

4.4. Efficiency Comparison

An efficiency comparison of our mixed randomized response technique and the Moors (1971) model by a variance comparison was done. From (4.3.2.12), we get

$$Var(\hat{\pi}_m)=\frac{\pi_S(1-\pi_S)}{n}+\frac{(1-P_1)[\lambda P_1(1-\pi_S)+(1-\lambda)]}{nP_1^2}.$$

From the optimization of unrelated question RR model, Moors (1971) derived the optimized Moors model:

$$Var(\hat{\pi}_{UM})=\frac{1}{nP_1^2}\left\{\sqrt{Y_1(1-Y_1)}+(1-P_1)\sqrt{\pi_I(1-\pi_I)}\right\}^2 \tag{4.4.1}$$

where $Y_1=P_1\pi_S+(1-P_1)\pi_I$.

We compute the relative efficiency, $Var(\hat{\pi}_{UM})/Var(\hat{\pi}_m)$, which is the proposed model based on estimator $\hat{\pi}_m$ with respect to the Moors (1990) model based on estimator $\hat{\pi}_{UM}$. The percent relative efficiency of $Var(\hat{\pi}_{UM})/Var(\hat{\pi}_m)$ is

$$\text{Percent RE}=\frac{\left\{\sqrt{Y_1(1-Y_1)}+(1-P_1)\sqrt{\pi_I(1-\pi_I)}\right\}^2}{P_1^2\pi_S(1-\pi_S)+(1-P_1)[\lambda P_1(1-\pi_S)+(1-\lambda)]}\times 100 \tag{4.4.2}$$

where $Y_1=P_1\pi_S+(1-P_1)\pi_I$.

If the percent RE is more than 100, then our proposed model is more efficient than the Moors (1971) model. Otherwise, the Moors model is more efficient than the proposed model. Since it is difficult to derive the mathematical condition of the relative efficiency from (4.3.2.12) and (4.4.1), an empirical investigation on the relative efficiency is presented in Table 4.1.

In Table 4.1, we allocated a sample size n to n_1 and n_2 by a way of estimating π_I since we asked a direct question about an innocuous trait to each respondent in a sample chosen from the population and deduced λ, which is the proportion of "Yes" answers to the direct question. Since n does not affect the computation of the percent RE, we did not change the sample size $n = 1000$.

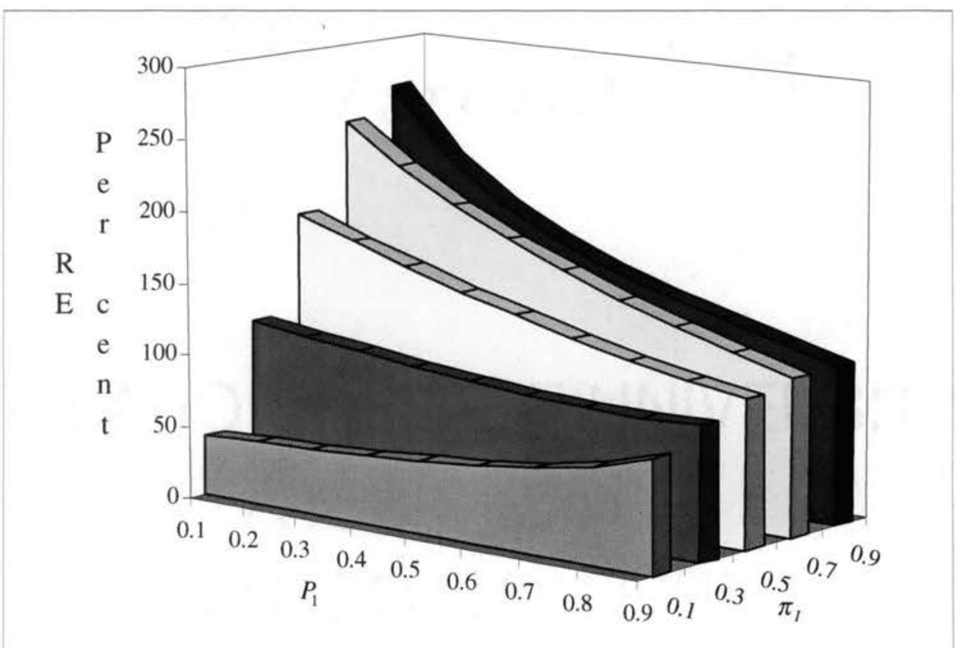

Figure 4.2. The Percent Relative Efficiency of $Var(\hat{\pi}_{UM}) / Var(\hat{\pi}_m)$ When $\pi_S = 0.2$.

TABLE 4.1.

The Percent Relative Efficiency of $Var(\hat{\pi}_{UM})/Var(\hat{\pi}_m)$.

		$n=1000$		The percent R.E.								
π_S	π_I	n_1	n_2	$P_1=.1$	$P_1=.2$	$P_1=.3$	$P_1=.4$	$P_1=.5$	$P_1=.6$	$P_1=.7$	$P_1=.8$	$P_1=.9$
0.1	0.9	900	100	261.3	216	188.3	168.8	153.6	141.1	130.1	120	110.1
	0.7	700	300	243.1	214.8	192.1	173.1	156.7	142.2	129.1	117.2	106.6
	0.5	500	500	183	167.9	154.2	141.7	130.3	119.8	110.3	102.1	96.45
	0.3	300	700	113.3	106.9	100.8	95.15	89.96	85.38	81.76	79.9	82.23
	0.1	100	900	39.67	39.51	39.59	40	40.91	42.61	45.68	51.43	63.68
0.2	0.9	900	100	266.5	220.7	192	171.3	155.3	142	130.4	119.9	109.9
	0.7	700	300	245.7	218.3	195.8	176.6	160	145.3	132	120	109.2
	0.5	500	500	184.7	170.8	158	146.3	135.5	125.6	116.7	108.9	102.8
	0.3	300	700	114.9	110	105.4	101.2	97.43	94.27	92	91.14	92.83
	0.1	100	900	41.49	43.18	45.14	47.49	50.42	54.21	59.37	66.84	78.63
0.3	0.9	900	100	272.4	226.2	196.3	174.6	157.5	143.3	131.1	120.1	109.8
	0.7	700	300	248.3	222.1	199.9	180.6	163.6	148.4	134.6	122.1	110.5
	0.5	500	500	186.5	173.8	161.9	150.8	140.4	130.6	121.6	113.3	106
	0.3	300	700	116.5	113	109.7	106.7	103.8	101.4	99.41	98.18	98.11
	0.1	100	900	43.25	46.59	50.1	53.91	58.17	63.09	68.98	76.36	86.13
0.4	0.9	900	100	278.9	232.7	201.7	178.7	160.4	145.3	132.2	120.6	110
	0.7	700	300	251.1	226.2	204.4	185	167.6	151.8	137.3	124	111.6
	0.5	500	500	188.2	176.9	165.9	155.3	145.1	135.3	125.9	116.8	108.2
	0.3	300	700	118	116	113.9	111.7	109.6	107.4	105.3	103.3	101.5
	0.1	100	900	44.95	49.78	54.62	59.57	64.74	70.26	76.27	83	90.74
0.5	0.9	900	100	286.2	240.5	208.4	184	164.4	148.1	134	121.6	110.4
	0.7	700	300	254.1	230.8	209.6	190.2	172.3	155.8	140.5	126.1	112.7
	0.5	500	500	190	180	170	160	150	140	130	120	110
	0.3	300	700	119.6	118.9	117.9	116.6	114.9	112.8	110.4	107.5	104
	0.1	100	900	46.59	52.79	58.78	64.65	70.47	76.3	82.15	88.05	94.01
0.6	0.9	900	100	294.6	250	217	191.1	170	152.1	136.8	123.2	111.1
	0.7	700	300	257.2	235.7	215.5	196.3	178.1	160.8	144.4	128.9	114.1
	0.5	500	500	191.8	183.3	174.3	165	155.2	145	134.4	123.4	111.9
	0.3	300	700	121.1	121.7	121.8	121.2	120	118	115.2	111.3	106.3
	0.1	100	900	48.19	55.65	62.65	69.28	75.59	81.57	87.16	92.24	96.61

		$n=1000$		The percent R.E.								
π_S	π_I	n_1	n_2	$P_1=.1$	$P_1=.2$	$P_1=.3$	$P_1=.4$	$P_1=.5$	$P_1=.6$	$P_1=.7$	$P_1=.8$	$P_1=.9$
	0.7	700	300	260.5	241.3	222.3	203.6	185.3	167.3	149.7	132.6	116
	0.5	500	500	193.7	186.7	178.9	170.4	161	150.8	139.6	127.4	114.2
	0.3	300	700	122.7	124.6	125.6	125.9	125.1	123.2	120	115.3	108.7
	0.1	100	900	49.74	58.38	66.28	73.58	80.29	86.36	91.67	96	98.97
0.8	0.9	900	100	315.2	277.1	244.2	215.6	190.5	168.2	148.3	130.5	114.5
	0.7	700	300	264	247.5	230.4	212.9	194.9	176.4	157.6	138.5	119.2
	0.5	500	500	195.6	190.2	183.9	176.4	167.8	157.7	146.2	132.8	117.5
	0.3	300	700	124.2	127.4	129.5	130.6	130.4	128.7	125.4	120	111.8
	0.1	100	900	51.25	60.99	69.73	77.62	84.7	90.89	96.01	99.74	101.5
0.9	0.9	900	100	328.2	297.6	268.1	240	213.2	187.7	163.5	140.9	119.7
	0.7	700	300	267.7	254.5	240.2	224.9	208.3	190.3	170.8	149.4	125.9
	0.5	500	500	197.5	194	189.3	183.4	176	166.9	155.6	141.6	123.7
	0.3	300	700	125.7	130.2	133.4	135.4	136.1	135.1	132.1	126.4	116.8
	0.1	100	900	52.72	63.51	73.02	81.48	88.95	95.35	100.5	103.9	104.7

We obtained the value of the percent relative efficiency for $\pi_I = 0.1, 0.3, 0.5, 0.7, 0.9$ and for different cases of π_S and P_1. From π_S, P_1 and $\pi_I \geq 0.5$ in Table 4.1, the value of the percent relative efficiency is more than 100 except for the case when $\pi_S = 0.1$. Furthermore, for $\pi_S \geq 0.4$ and $\pi_I \geq 0.3$, the value of the percent relative efficiency is more than 100 for all values of P_1. Figure 4.2 shows that our mixed model is always more efficient than the Moors model if the "Yes" proportion from the innocuous trait direct is more than half percent. Under the condition $\lambda = (n_1/n) > 0.5$, we can conclude that the mixed randomized response model is a good alternative strategy of the Moors model while keeping the confidentiality of interviewee.

4.5.A Mixed Randomized Response Model Using Stratification

4.5.1. A mixed stratified RR model

A stratified Warner's randomized response model was presented in Chapter III. So we apply a stratified randomized response technique to the proposed mixed model. The assumptions for a stratified mixed model are the same as those in Chapter III. Thus, the population is partitioned into k strata, and a sample is selected by simple random sampling with replacement within each stratum.

Assume that the number of units from each stratum is known. An individual respondent in a sample of each stratum is instructed to answer a direct statement "I am a member of the innocuous trait group". Respondents should answer the direct statement by "Yes" or "No".

If a respondent answers "Yes", then she or he is instructed to go to a randomization device S_{h1} consisting of the two statements. The one statement is "I am a member of the sensitive trait group", and the other one is "I am a member of the innocuous trait group" with preassigned probabilities, Q_h and $1-Q_h$ respectively.

If a respondent answers "No", then the respondent is instructed to go to a randomization device S_{h2} consisting of the two statements. The one statement is "I am a member of a sensitive trait group", and the other is "I am not a member of a sensitive trait group" with preassigned probabilities, P_h and $1-P_h$ respectively.

To protect respondents' privacy, respondents should not disclose the question they had from S_{h1} or S_{h2} to the interviewer. Suppose we denote m_h as the number of units in the sample from stratum h and n as the total number of units in samples from all strata.

Let m_{h1} be the number of people responding "Yes" when respondents in a sample m_h were asked the direct question and m_{h2} be the number of people responding "No" when respondents in a sample m_h were asked the direct question such that $n = \sum_{h=1}^{k} (m_{h1} + m_{h2})$. Under this assumption that these "Yes" or "No" reports are made truthfully, and Q_h and $P_h (\neq 0.5)$ are set by a researcher, then the proportion of "Yes" answer from the respondents using a randomization device S_{h1} is

$$Y_{1h} = Q_h \pi_{S_h} + (1 - Q_h) \pi_{I_h} \qquad \text{for } h = 1, 2, \cdots, k. \qquad (4.5.1)$$

where Y_{1h} is the proportion of "Yes" answer in a stratum h, π_{S_h} is the proportion of respondents with the sensitive trait in a stratum h, π_{I_h} is the proportion of respondents with the innocuous trait in a stratum h, and Q_h is the probability that a respondent in the sample stratum h is asked a sensitive question.

Since the respondent performing a randomization device R_1 responded "Yes" to the direct question about the innocuous question, π_{I_h} is equal to one.

(4.5.1) becomes $Y_{1h} = Q_h \pi_{S_h} + (1 - Q_h)$. The estimator of π_{S_h} is

$$\hat{\pi}_{UY_h} = \frac{\hat{Y}_{1h} - (1 - Q_h)}{Q_h} \qquad \text{for } h = 1, 2, \cdots, k. \qquad (4.5.2)$$

where $\hat{Y}_{1h}$ is the proportion of "Yes" answer in a sample in the stratum h and $\hat{\pi}_{UY_h}$ is the proportion of respondents with the sensitive trait in a sample from the stratum h.

Since each $\hat{Y}_{1h}$ is a binomial distribution, $B(m_{h1}, Y_{1h})$, the estimator $\hat{\pi}_{UK_h}$ is unbiased for π_{S_h} with

$$Var(\hat{\pi}_{UY_h}) = \frac{Y_{1h}(1-Y_{1h})}{m_{h1}Q_h^{\ 2}} = \frac{[Q_h\pi_{S_h} + (1-Q_h)][1-Q_h\pi_{S_h} - (1-Q_h)]}{m_{h1}Q_h^{\ 2}}.$$

$$= \frac{Q_h(1-\pi_{S_h})[Q_h\pi_{S_h} + (1-Q_h)]}{m_{h1}Q_h^{\ 2}} = \frac{(1-\pi_{S_h})[Q_h\pi_{S_h} + (1-Q_h)]}{m_{h1}Q_h}. \tag{4.5.3}$$

The proportion of "Yes" answer from the respondents performing a randomization device S_{h2} is

$$X_h = P_h\pi_{S_h} + (1-P_h)(1-\pi_{S_h}) = (2P_h - 1)\pi_{S_h} + (1-P_h) \quad \text{for } h = 1,2,\cdots,k \tag{4.5.4}$$

where X_h is the proportion of "Yes" answer in a stratum h, π_{S_h} is the proportion of respondents with the sensitive trait in a stratum h, and P_h is the probability that a respondent in the sample stratum h has a sensitive question card.

The maximum likelihood estimate in this case is

$$\hat{\pi}_{W_h} = \frac{\hat{X}_h - (1-P_h)}{2P_h - 1} \qquad \text{for } h = 1,2,\cdots,k \tag{4.5.5}$$

where $\hat{X}_h$ is the proportion of "Yes" answer in a sample from stratum h and $\hat{\pi}_{W_h}$ is the proportion of respondents with the sensitive trait in a sample from stratum h.

Since each $\hat{X}_h$ is a binomial distribution $B(m_h, X_h)$, the estimator $\hat{\pi}_{WS_h}$ is unbiased for π_{S_h}. For its variance,

$$Var(\hat{\pi}_{W_h}) = \frac{\pi_{S_h}(1-\pi_{S_h})}{m_{h2}} + \frac{P_h(1-P_h)}{m_{h2}(2P_h - 1)^2}. \tag{4.5.6}$$

By the theorem 4.3.2.1,

$$Var(\hat{\pi}_{W_h}) = \frac{\pi_{S_h}(1-\pi_{S_h})}{m_{h2}} + \frac{1-Q_h}{m_{h2}Q_h^{\;2}}. \tag{4.5.7}$$

Then the estimator of π_{S_h}, in terms of sample proportions of "Yes" responses, $\hat{Y}_{1h}$ and $\hat{X}_h$, as

$$\hat{\pi}_{mS_h} = \frac{m_{h1}}{m_h}\hat{\pi}_{UY_h} + \frac{m_{h2}}{m_h}\hat{\pi}_{W_h} \qquad \text{for } 0 < \frac{m_{h1}}{m_h} \text{ and } \frac{m_{h2}}{m_h} < 1 \tag{4.5.8}$$

By the theorem 4.3.2.1, the proposed estimator $\hat{\pi}_{mS_h}$ is unbiased for population proportion π_{S_h}. By the theorem 4.3.2.3,

$$Var(\hat{\pi}_{mS_h}) = Var\left(\frac{m_{h1}}{m_h}\hat{\pi}_{UY_h} + \frac{m_{h2}}{m_h}\hat{\pi}_{W_h}\right) = \left(\frac{m_{h1}}{m_h}\right)^2 Var(\hat{\pi}_{UY_h}) + \left(\frac{m_{h2}}{m_h}\right)^2 Var(\hat{\pi}_{W_h})$$

$$= \frac{\pi_{S_h}(1-\pi_{S_h})}{m_h} + \frac{(1-Q_h)[\lambda_h Q_h(1-\pi_{S_h}) + (1-\lambda_h)]}{m_h Q_h^{\;2}} \tag{4.5.9}$$

where $m_h = m_{h1} + m_{h2}$ and $\lambda_h = m_{h1}/m_h$.

By the theorem 4.3.2.4, an unbiased estimate of $Var(\hat{\pi}_{mS_h})$ is given by

$$v(\hat{\pi}_{mS_h}) = \frac{\hat{\pi}_{mS_h}(1-\hat{\pi}_{mS_h})}{m_h^{\;2}}\left[\frac{m_{h1}^{\;2}}{m_{h1}-1} + \frac{m_{h2}^{\;2}}{m_{h2}-1}\right] + \frac{(1-Q_h)}{(m_h Q_h)^2}\left[\frac{m_{h1}^{\;2} Q_h(1-\hat{\pi}_{mS_h})}{m_{h1}-1} + \frac{m_{h2}^{\;2}}{m_{h2}-1}\right]. \tag{4.5.10}$$

Since the selections in different strata are made independently, the estimators for individual strata can be added together to obtain an estimator for the whole population.

The estimator of π_S is shown to be

$$\hat{\pi}_{mS} = \sum_{h=1}^{k} w_h \hat{\pi}_{mS_h} = \sum_{h=1}^{k} w_h \left[\frac{m_{h1}}{m_h}\hat{\pi}_{UY_h} + \frac{m_{h2}}{m_h}\hat{\pi}_{W_h}\right] \tag{4.5.11}$$

where N is the number of units in the whole population, N_h is the total number of units in the stratum h and $w_h = \frac{N_h}{N}$ for $h = 1,2,\cdots,k$ so that $w = \sum_{h=1}^{k} w_h = 1$.

Theorem 4.5.1. The proposed estimator $\hat{\pi}_{mS}$ is unbiased for the sensitive proportion π_S of the population.

Proof. As each estimator $\hat{\pi}_{mS_h}$ is unbiased for π_{S_h}, the expected value of $\hat{\pi}_{mS}$ is

$$E(\hat{\pi}_{mS}) = E\left(\sum_{h=1}^{k} w_h \hat{\pi}_{mS_h}\right) = \sum_{h=1}^{k} w_h E(\hat{\pi}_{mS_h}) = \pi_S .$$

An estimator $\hat{\pi}_{mS}$ of π_S is unbiased.

Theorem 4.5.2. The variance of an estimator $\hat{\pi}_{mS}$ is given by

$$Var(\hat{\pi}_{mS}) = \sum_{h=1}^{k} \frac{w_h^{\,2}}{m_h}\left[\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h}) + (1-\lambda_h)\}}{Q_h^{\,2}}\right] \qquad (4.5.12)$$

where $m_h = m_{h1} + m_{h2}$ and $\lambda_h = m_{h1}/m_h$.

Proof. Since each unbiased estimator $\hat{\pi}_{mS_h}$ has its own variance and strata are independent, the variance of $\hat{\pi}_{mS}$ using (4.5.9) and Corollary 1. in Sec. 5.9 of Cochran (1977) is

$$Var(\hat{\pi}_{mS}) = Var\left(\sum_{h=1}^{k} w_h \hat{\pi}_{mS_h}\right) = \sum_{h=1}^{k} w_h^{\,2} Var(\hat{\pi}_{mS_h})$$

$$= \sum_{h=1}^{k} w_h^{\,2}\left[\frac{\pi_{S_h}(1-\pi_{S_h})}{m_h} + \frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h}) + (1-\lambda_h)\}}{m_h Q_h^{\,2}}\right]$$

$$= \sum_{h=1}^{k} \frac{w_h^{\,2}}{m_h}\left[\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h}) + (1-\lambda_h)\}}{Q_h^{\,2}}\right]$$

which proves the theorem.

Theorem 4.5.3. The unbiased estimator of the variance $Var(\hat{\pi}_{mS})$ is given by

$$v(\hat{\pi}_{mS}) = \sum_{h=1}^{k} \frac{w_h^2}{m_h^2} \left[\hat{\pi}_{mS_h}(1-\hat{\pi}_{mS_h}) \left\{ \frac{m_{h1}^2}{m_{h1}-1} + \frac{m_{h2}^2}{m_{h2}-1} \right\} + \frac{(1-Q_h)}{Q_h^2} \left\{ \frac{m_{h1}^2 Q_h (1-\hat{\pi}_{mS_h})}{m_{h1}-1} + \frac{m_{h2}^2}{m_{h2}-1} \right\} \right]$$
(4.5.13) .

Proof. The proof is similar to theorem 4.5.2.

In order to do the optimal allocation of a sample size n, we need to know $\lambda_h = m_{h1}/m_h$ and π_{S_h}. Information on $\lambda_h = m_{h1}/m_h$ and π_{S_h} is usually unavailable. But if prior information about them is available from past experience or a pilot survey then it helps to derive the following optimal allocation formula.

Theorem 4.5.4. The optimal allocation n to $m_1, m_2, \cdots, m_{k-1}$ and m_k to derive the minimal value of variance of an estimator $\hat{\pi}_{mS}$ subject to $n = \sum_{h=1}^{k} m_h$ is approximately given by

$$\frac{m_h}{n} = \frac{w_h \left[\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-Q_h)\{\lambda_h Q_h (1-\pi_{S_h}) + (1-\lambda_h)\}}{Q_h^2} \right]^{1/2}}{\sum_{h=1}^{k} w_h \left[\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-Q_h)\{\lambda_h Q_h (1-\pi_{S_h}) + (1-\lambda_h)\}}{Q_h^2} \right]^{1/2}} \qquad (4.5.14)$$

where $m_h = m_{h1} + m_{h2}$ and $\lambda_h = m_{h1}/m_h$.

Proof. Suppose that $m_h = m_{h1} + m_{h2}$ and $\lambda_h = m_{h1}/m_h$. For minimum variance for fixed total sample size in Sec. 5.9 of Cochran (1977),

$$m_h \propto N_h \left[\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-Q_h)\{\lambda_h Q_h (1-\pi_{S_h}) + (1-\lambda_h)\}}{Q_h^2} \right]^{1/2} .$$

Thus

$$\frac{m_h}{n} \doteq \frac{N_h\left[\pi_{S_h}(1-\pi_{S_h})+\dfrac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^{\ 2}}\right]^{1/2}}{\sum_{h=1}^{k} N_h\left[\pi_{S_h}(1-\pi_{S_h})+\dfrac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^{\ 2}}\right]^{1/2}}$$

$$= \frac{w_h\left[\pi_{S_h}(1-\pi_{S_h})+\dfrac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^{\ 2}}\right]^{1/2}}{\sum_{h=1}^{k} w_h\left[\pi_{S_h}(1-\pi_{S_h})+\dfrac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^{\ 2}}\right]^{1/2}}.$$

The portion of the total sample size which should be allocated to each sample is (4.5.14).

Corollary 4.1. If we insert (4.5.13) into the following inequality:

$$\left[\sum_{h=1}^{k}\frac{w_h^{\ 2}}{m_h}\left\{\pi_{S_h}(1-\pi_{S_h})+\frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^{\ 2}}\right\}\right]\left(\sum_{i=1}^{k} m_i\right)$$

$$\geq\left[\sum_{h=1}^{k} w_h\left\{\pi_{S_h}(1-\pi_{S_h})+\frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^{\ 2}}\right\}^{1/2}\right]^2 \qquad (4.5.15)$$

then we can easily show

$$\left[\sum_{h=1}^{k}\frac{w_h^{\ 2}}{m_h}\left\{\pi_{S_h}(1-\pi_{S_h})+\frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^{\ 2}}\right\}\right]\left(\sum_{i=1}^{k} m_i\right)$$

$$=\left[\sum_{h=1}^{k} w_h\left\{\pi_{S_h}(1-\pi_{S_h})+\frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^{\ 2}}\right\}^{1/2}\right]^2 \qquad (4.5.16)$$

Using (4.5.16), we derive the minimal variance of an estimator $\hat{\pi}_{mS}$ in the following theorem.

Theorem 4.5.5. The minimal variance of the estimator $\hat{\pi}_{mS}$ is given by

$$Var(\hat{\pi}_{mS}) = \frac{1}{n}\left[\sum_{h=1}^{k} w_h \left\{\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h}) + (1-\lambda_h)\}}{Q_h^{\,2}}\right\}^{1/2}\right]^2 \quad (4.5.17)$$

where $n = \sum_{h=1}^{k} m_h$, $m_h = m_{h1} + m_{h2}$ and $\lambda_h = m_{h1}/m_h$.

Proof. By using (4.5.12), (4.5.14) and (4.5.16), $Var(\hat{\pi}_{mS})$ reduces to (4.5.17).

4.5.2. An Efficiency Comparison of a Stratified RR Model

We will do an efficiency comparison of a stratified mixed randomized response technique and the mixed randomized response model by comparing $Var(\hat{\pi}_{mS})$ and $Var(\hat{\pi}_m)$: From (4.5.17), we get

$$Var(\hat{\pi}_{mS}) = \frac{1}{n}\left[\sum_{h=1}^{k} w_h \left\{\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h}) + (1-\lambda_h)\}}{Q_h^{\,2}}\right\}^{1/2}\right]^2$$

where $n = \sum_{h=1}^{k} m_h$, $m_h = m_{h1} + m_{h2}$ and $\lambda_h = m_{h1}/m_h$.

From (4.3.2.12), we get the variance of $\hat{\pi}_m$: For $n = n_1 + n_2$ and $\lambda = n_1/n$,

$$Var(\hat{\pi}_m) = \frac{\pi_S(1-\pi_S)}{n} + \frac{(1-P_1)[\lambda P_1(1-\pi_S) + (1-\lambda)]}{nP_1^2}.$$

The following theorem is an efficiency comparison of a stratified mixed randomized response model and a mixed randomized response model.

Theorem 4.5.6. Suppose there are two strata in the population and $\lambda_h = m_{h1}/m_h$. The $\hat{\pi}_{mS}$ of a stratified mixed RR is more efficient than the estimator $\hat{\pi}_m$ of a mixed model where $P_1 = Q_1 = Q_2$ and $\lambda = \lambda_1 = \lambda_2$.

Proof. Assume $n = n_1 + n_2$, $\hat{\pi}_S = w_1\hat{\pi}_{S_1} + w_2\hat{\pi}_{S_2}$ and $P_1 = Q_1 = Q_2$ and $\lambda = \lambda_1 = \lambda_2$. Using (4.5.17) and (4.3.2.12), the efficiency of $\hat{\pi}_{mS}$ with respect to $\hat{\pi}_m$ is given by

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_{mS}) = \frac{\pi_S(1-\pi_S)}{n} + \frac{(1-P_1)[\lambda P_1(1-\pi_S)+(1-\lambda)]}{nP_1^2}$$

$$-\frac{1}{n}\left[\sum_{h=1}^{2} w_h\left\{\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-P_1)\{\lambda P_1(1-\pi_{S_h})+(1-\lambda)\}}{P_1^2}\right\}^{1/2}\right]^2.$$

Inserting $\pi_S = w_1\pi_{S_1} + w_2\pi_{S_2}$ such that $\pi_{S_1} \neq \pi_{S_2}$ into the above equation, then we can derive:

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_{mS}) = \frac{(w_1\pi_{S_1} + w_2\pi_{S_2}) - (w_1^2\pi_{S_1}^2 + w_2^2\pi_{S_2}^2) - 2w_1w_2\pi_{S_1}\pi_{S_2}}{n}$$

$$+\frac{(1-P_1)[\lambda P_1(1-\pi_S)+(1-\lambda)]}{nP_1^2}$$

$$-\frac{1}{n}\left[\sum_{h=1}^{2} w_h\left\{\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-P_1)\{\lambda P_1(1-\pi_{S_h})+(1-\lambda)\}}{P_1^2}\right\}^{1/2}\right]^2.$$

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_{mS}) = \frac{w_1\pi_{S_1} + w_2\pi_{S_2} - 2w_1w_2\pi_{S_1}\pi_{S_2}}{n} + \frac{(1-P_1)[\lambda P_1(1-\pi_S)+(1-\lambda)]}{nP_1^2}$$

$$-\left(\frac{w_1^2\pi_{S_1} + w_2^2\pi_{S_2}}{n}\right) - \frac{1}{n}\left[\sum_{h=1}^{2} w_h^2\left\{\frac{(1-P_1)\{\lambda P_1(1-\pi_{S_h})+(1-\lambda)\}}{P_1^2}\right\}\right]$$

$$-\frac{2w_1(1-w_1)}{n}\prod_{h=1}^{2}\left[\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-P_1)\{\lambda P_1(1-\pi_{S_h})+(1-\lambda)\}}{P_1^2}\right]^{1/2}.$$

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_{mS}) = \frac{w_1 w_2 (\pi_{S_1} + \pi_{S_2} - 2\pi_{S_1}\pi_{S_2})}{n} + \frac{\lambda(1-P_1)(1-\pi_S)}{nP_1} + \frac{(1-P_1)(1-\lambda)}{nP_1^2}$$

$$-\frac{1}{n}\left[\sum_{h=1}^{2} w_h^{\ 2}\left\{\frac{\lambda(1-P_1)(1-\pi_{S_h})}{P_1}\right\}\right] - \frac{1}{n}\left[\sum_{h=1}^{2} w_h^{\ 2}\left\{\frac{(1-P_1)(1-\lambda)}{P_1^2}\right\}\right]$$

$$-\frac{2w_1 w_2}{n}\prod_{h=1}^{2}\left[\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-P_1)\{\lambda P_1(1-\pi_{S_h}) + (1-\lambda)\}}{P_1^2}\right]^{1/2}.$$

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_{mS}) = \frac{w_1 w_2 (\pi_{S_1} + \pi_{S_2} - 2\pi_{S_1}\pi_{S_2})}{n} + \frac{\lambda(1-P_1)(1-w_1\pi_{S_1} - w_2\pi_{S_2})}{nP_1}$$

$$-\frac{1}{n}\left[\sum_{h=1}^{2} w_h^{\ 2}\left\{\frac{\lambda(1-P_1)(1-\pi_{S_h})}{P_1}\right\}\right] + \frac{(1-w_1^{\ 2} - w_2^{\ 2})(1-P_1)(1-\lambda)}{nP_1^2}$$

$$-\frac{2w_1 w_2}{n}\prod_{h=1}^{2}\left[\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-P_1)\{\lambda P_1(1-\pi_{S_h}) + (1-\lambda)\}}{P_1^2}\right]^{1/2}.$$

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_{mS}) = \frac{w_1 w_2 (\pi_{S_1} + \pi_{S_2} - 2\pi_{S_1}\pi_{S_2})}{n} + \frac{2w_1 w_2 (1-P_1)(1-\lambda)}{nP_1^2}$$

$$+\frac{\lambda(1-P_1)}{nP_1}\left[1 - w_1\pi_{S_1} - w_2\pi_{S_2} - w_1^{\ 2}(1-\pi_{S_1}) - w_2^{\ 2}(1-\pi_{S_2})\right]$$

$$-\frac{2w_1 w_2}{n}\prod_{h=1}^{2}\left[\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-P_1)\{\lambda P_1(1-\pi_{S_h}) + (1-\lambda)\}}{P_1^2}\right]^{1/2}.$$

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_{mS}) = \frac{w_1 w_2}{n}\left[\pi_{S_1}(1-\pi_{S_2}) + \pi_{S_2}(1-\pi_{S_1}) + \frac{2(1-P_1)(1-\lambda)}{P_1^2}\right.$$

$$\left. + \frac{\lambda(1-P_1)(2-\pi_{S_1} - \pi_{S_2})}{P_1}\right] - \frac{2w_1 w_2}{n}\prod_{h=1}^{2}\left[\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-P_1)\{\lambda P_1(1-\pi_{S_h}) + (1-\lambda)\}}{P_1^2}\right]^{1/2}.$$

Let $A = \pi_{S_1}(1-\pi_{S_1}) + \dfrac{(1-P_1)\{\lambda P_1(1-\pi_{S_1}) + (1-\lambda)\}}{P_1^2}$ and

$$B = \pi_{S_2}(1-\pi_{S_2}) + \frac{(1-P_1)\{\lambda P_1(1-\pi_{S_2}) + (1-\lambda)\}}{P_1^2}.$$

Then we can derive the following one:

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_{mS}) = \frac{w_1 w_2}{n}\left[\pi_{S_1}(1-\pi_{S_2}) + \pi_{S_2}(1-\pi_{S_1}) + \frac{2(1-P_1)(1-\lambda)}{P_1^2}\right.$$
$$\left. + \frac{\lambda(1-P_1)(2-\pi_{S_1}-\pi_{S_2})}{P_1}\right] + \frac{w_1 w_2}{n}\left[\left(\sqrt{A}-\sqrt{B}\right)^2 - A - B\right].$$

Hence

$$Var(\hat{\pi}_m) - Var(\hat{\pi}_{mS}) = \frac{w_1 w_2}{n}\left(\sqrt{A}-\sqrt{B}\right)^2 > 0 \qquad \text{where } \pi_{S_1} \neq \pi_{S_2}.$$

Since $Var(\hat{\pi}_{ms}) - Var(\hat{\pi}_S) > 0$, then the estimator $\hat{\pi}_{mS}$ of a stratified mixed RR is more efficient than the estimator $\hat{\pi}_m$ of a mixed model.

In Table 4.1, we showed that the mixed RR model is more efficient than the Moors (1971) model where $\pi_I > 0.5$. We will derive the mathematical condition for the efficiency of a stratified mixed RR model and the Moors model. From (4.4.1), we get the optimized Moors model:

For $Y_1 = P_1\pi_S + (1-P_1)\pi_I$,

$$Var(\hat{\pi}_{UM}) = \frac{1}{nP_1^2}\left\{\sqrt{Y_1(1-Y_1)} + (1-P_1)\sqrt{\pi_I(1-\pi_I)}\right\}^2.$$

Theorem 4.5.7. The estimator $\hat{\pi}_{mS}$ of a stratified mixed RR model is more efficient than the estimator $\hat{\pi}_{UM}$ of the Moors model if

$$\frac{\sqrt{Y_1(1-Y_1)}+\sqrt{\pi_I(1-\pi_I)}}{\sqrt{\pi_I(1-\pi_I)}+\sum_{h=1}^{k} w_h \sqrt{\pi_{S_h}(1-\pi_{S_h})+\frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^2}}} > P_1 \qquad (4.5.18)$$

where $\lambda_h = m_{h1}/m_h$ and $Y_1 = P_1\pi_S + (1-P_1)\pi_I$.

Proof. Assume $\lambda_h = m_{h1}/m_h$ and $Y_1 = P_1\pi_S + (1-P_1)\pi_I$. Using (4.5.17) and (4.4.1), we check the efficiency of $\hat{\pi}_{mS}$ with respect to $\hat{\pi}_{UM}$.

$$Var(\hat{\pi}_{UM}) - Var(\hat{\pi}_{mS}) = \frac{1}{nP_1^2}\left\{\sqrt{Y_1(1-Y_1)}+(1-P_1)\sqrt{\pi_I(1-\pi_I)}\right\}^2$$

$$-\frac{1}{n}\left[\sum_{h=1}^{k} w_h \left\{\pi_{S_h}(1-\pi_{S_h})+\frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^2}\right\}^{1/2}\right]^2.$$

$$Var(\hat{\pi}_{UM}) - Var(\hat{\pi}_{mS}) = \frac{1}{n}\left\{\frac{\sqrt{Y_1(1-Y_1)}+(1-P_1)\sqrt{\pi_I(1-\pi_I)}}{P_1}\right\}^2$$

$$-\frac{1}{n}\left[\sum_{h=1}^{k} w_h \left\{\pi_{S_h}(1-\pi_{S_h})+\frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^2}\right\}^{1/2}\right]^2.$$

Let $L = \dfrac{\sqrt{Y_1(1-Y_1)}+(1-P_1)\sqrt{\pi_I(1-\pi_I)}}{P_1}$

and $M = \sqrt{\pi_{S_h}(1-\pi_{S_h})+\dfrac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h})+(1-\lambda_h)\}}{Q_h^2}}$.

Then

$$Var(\hat{\pi}_{UM}) - Var(\hat{\pi}_{mS}) = \frac{1}{n}\left(L^2 - M^2\right) = \frac{1}{n}(L+M)(L-M).$$

If $Var(\hat{\pi}_{UM}) - Var(\hat{\pi}_{mS}) > 0$, then the estimator $\hat{\pi}_{mS}$ of a stratified mixed RR is more efficient than the estimator $\hat{\pi}_{UM}$ of the Moors model. Since $L+M$ is positive, if $L-M > 0$ then $Var(\hat{\pi}_{UM}) - Var(\hat{\pi}_{mS}) > 0$. Suppose $L-M > 0$.

$$\frac{\sqrt{Y_1(1-Y_1)} + (1-P_1)\sqrt{\pi_I(1-\pi_I)}}{P_1} > \sqrt{\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h}) + (1-\lambda_h)\}}{Q_h^{\ 2}}}.$$

Hence,

$$\sqrt{Y_1(1-Y_1)} + \sqrt{\pi_I(1-\pi_I)}$$

$$> P_1\left[\sqrt{\pi_I(1-\pi_I)} + \sqrt{\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h}) + (1-\lambda_h)\}}{Q_h^{\ 2}}}\right].$$

If prior information on π_S, π_I, w_h, π_{S_h}, Q_h, P_1 and λ_h satisfy the following condition:

$$\frac{\sqrt{Y_1(1-Y_1)} + \sqrt{\pi_I(1-\pi_I)}}{\sqrt{\pi_I(1-\pi_I)} + \sum_{h=1}^{k} w_h \sqrt{\pi_{S_h}(1-\pi_{S_h}) + \frac{(1-Q_h)\{\lambda_h Q_h(1-\pi_{S_h}) + (1-\lambda_h)\}}{Q_h^{\ 2}}}} > P_1$$

then the estimator $\hat{\pi}_{mS}$ of a stratified mixed RR is more efficient than the estimator $\hat{\pi}_{UM}$ of the Moors model.

Suppose that prior information on π_{S_1}, π_{S_2}, w_1, w_2, π_S, π_I, w_h, π_{S_h} and λ_h can be roughly obtained and a researcher sets P_1 and Q_h. Then we can do the efficiency comparison of a mixed stratified RR model and the Moors model (1971) under condition (4.5.18). In this paper, we use the percent relative efficiency of $Var(\hat{\pi}_{UM}) / Var(\hat{\pi}_{mS})$ which is a stratified mixed model based on estimator $\hat{\pi}_{mS}$ with respect to the Moors (1990) model based on estimator $\hat{\pi}_{UM}$.

The percent relative efficiency is

$$\text{Percent RE} = \frac{Var(\hat{\pi}_{UM})}{Var(\hat{\pi}_{mS})} \times 100\,. \tag{4.5.19}$$

Hence, if the percent RE is more than 100, then our proposed model is more efficient than the Moors (1971) model. Otherwise, that is, if the percent RE is less than or equal to 100 then the Moors model is more efficient than the proposed model. Values of the percent RE for different sets of prior information are presented in the Table 4.2. Since λ_h is the proportion of "Yes" for the direct question of an innocuous attribute in the stratum h, we set $\pi_I = \lambda_1 = \lambda_2$ and $P_1 = Q_1$ for the convenience of the efficiency comparison in the Table 4.1. For fixed $w_1 = 0.6$ and $w_2 = 0.4$, we changed the values of π_{S_1} and π_{S_2} and increased the value of π_I from 0.2 to 0.8 by 0.2 increments. We did not change the sample size $n = 1000$ in Table 4.2 because n does not affect the computation of the percent RE. For different cases of $P_1 = Q_1$ and Q_2, we compared the efficiency of a stratified mixed RR model and the Moors (1971) model. The observation in Table 4.2 is that, under the condition (4.5.18), the values of the percent relative efficiency are more than 100. This shows that the estimator $\hat{\pi}_{mS}$ of a stratified mixed RR model is more efficient than the estimator $\hat{\pi}_{UM}$ of the Moors model.

Figure 4.3 shows the result from Table 4.2 when $P_1 = Q_1 = 0.5$ and $Q_2 = 0.8$. We examined the relative efficiency for a case with two strata. For a case with more than two strata, we may derive similar conclusions in terms of efficiency as those for two strata.

TABLE 4.2.

The Percent Relative Efficiency of $Var(\hat{\pi}_{UM})/Var(\hat{\pi}_{mS})$ when $n=1000$.

						$P_1=Q_1$							
						0.1		0.2		0.3		0.4	
						Q_2		Q_2		Q_2		Q_2	
π_{S_1}	π_{S_2}	w_1	w_2	π_S	π_I	0.2	0.3	0.3	0.4	0.5	0.6	0.7	0.8
0.08	0.13	0.6	0.4	0.1	0.2	122.7	147.7	101.3	121.5	107.7	122	112.9	125.6
					0.4	235.8	283	188.3	224.6	192.5	217.1	193.6	215.1
					0.6	335.7	400.9	259.8	307.5	256.8	288.4	251.8	279.2
					0.8	397.5	470.2	293.8	343.3	280	312.6	270	298.9
0.18	0.23	0.6	0.4	0.2	0.2	125.4	150.8	105.8	126.7	114.9	129.5	122.1	134.8
					0.4	238.5	286	192.2	229	197.6	222	198.6	218.6
					0.6	339.3	404.9	263.9	311.9	260.7	291.4	253.2	277.9
					0.8	404	477.5	299.2	348.8	283.3	314.4	268.6	293.8
0.28	0.33	0.6	0.4	0.3	0.2	128	154	110.2	131.8	121.5	136.7	130.5	143.3
					0.4	241.2	289.2	196.2	233.5	202.9	227.4	204.1	223.3
					0.6	343.1	409.5	268.4	316.9	265.5	295.7	256.3	279.3
					0.8	411.2	485.9	305.5	355.7	288	318.2	269.6	292.5
0.38	0.43	0.6	0.4	0.4	0.2	130.7	157.1	114.4	136.8	127.9	143.7	138.4	151.5
					0.4	243.9	292.6	200.3	238.3	208.6	233.4	210.3	229.3
					0.6	347.2	414.4	273.4	322.8	271.2	301.6	261.1	283.3
					0.8	419.2	495.6	313	364.2	294.5	324.4	273.1	294.5
0.48	0.53	0.6	0.4	0.5	0.2	133.3	160.3	118.5	141.8	134.1	150.6	146.3	160
					0.4	246.8	296.1	204.5	243.5	214.7	240.1	217.4	236.7
					0.6	351.6	419.9	278.9	329.6	278.2	309.1	267.9	290
					0.8	428.2	506.7	321.9	374.9	303.2	333.4	279.5	300.3
0.58	0.63	0.6	0.4	0.6	0.2	135.9	163.5	122.6	146.7	140.3	157.7	154.3	169
					0.4	249.8	299.9	209	249.1	221.4	247.9	225.8	246
					0.6	356.3	426	285.1	337.5	286.7	318.8	277.3	300.1
					0.8	438.4	519.5	332.6	388.1	314.9	346.2	289.7	310.6
0.68	0.73	0.6	0.4	0.7	0.2	138.4	166.7	126.6	151.7	146.5	165.1	163	179.1
					0.4	252.9	303.9	213.7	255.2	229	256.9	236.1	258
					0.6	361.3	432.8	292.1	346.7	297.2	331.2	290.2	314.8
					0.8	449.8	534.5	345.7	404.8	330.8	364.3	305.4	327.5
0.78	0.83	0.6	0.4	0.8	0.2	141	169.9	130.6	156.8	153.1	173	172.7	190.9
					0.4	256	308.1	218.6	261.8	237.5	267.6	249	273.9
					0.6	366.7	440.3	300	357.6	310.2	347.5	308.3	336.5
					0.8	462.8	552.1	361.9	426.4	353	390.7	330.4	355.8
0.88	0.93	0.6	0.4	0.9	0.2	143.6	173.2	134.6	162	159.9	181.7	184	205.7
					0.4	259.4	312.7	223.9	269.2	247.4	280.5	266	296.2
					0.6	372.6	448.6	309	370.5	326.9	369.3	335.2	370.8
					0.8	477.8	573.1	382.3	455	385.9	431.8	374.3	408.4

		$P_1 = Q_1$									
		0.5		0.6		0.7		0.8		0.9	
		Q_2		Q_2		Q_2		Q_2		Q_2	
π_S	π_I	0.8	0.9	0.86	0.9	0.95	0.96	0.95	0.96	0.95	0.96
0.1	0.2	107.9	121.7	100.8	106.9	102.4	104.3	**	**	**	**
	0.4	177.5	200	158.2	167.5	151.2	153.9	125.3	128	102.5	105.2
	0.6	226.6	255	198.4	209.9	185.5	188.8	149.3	152.4	116	119
	0.8	242.2	272.3	212.6	224.7	199.3	202.7	160.8	164.1	123.8	126.9
0.2	0.2	118.5	131.6	112.5	118.1	113.8	115.4	103.8	105.5	**	**
	0.4	182	201.7	162.5	170.4	153.1	155.1	129	131	108	110
	0.6	225.6	249.3	196.1	205.3	178.9	181.2	145.7	147.9	116.5	118.5
	0.8	237.1	261.2	205.2	214.5	186.5	188.8	151.6	153.7	120.3	122.2
0.3	0.2	127.5	140.4	121.6	127.1	122	123.4	111	112.5	100.9	102.5
	0.4	187	205.2	166.9	174.1	155.7	157.5	131.8	133.5	110.7	112.4
	0.6	226.9	248	196.1	204.1	176.5	178.4	144.4	146.1	116.8	118.4
	0.8	235.2	255.8	201.6	209.3	180	181.8	147.1	148.7	118.6	120.1
0.4	0.2	135.9	148.9	129.7	135.2	129	130.4	116.5	117.9	104.4	105.8
	0.4	192.8	210.4	172	178.9	159.3	160.9	134.6	136.1	112.7	114.2
	0.6	230.3	249.9	198.2	205.5	176.7	178.3	144.7	146.2	117.4	118.8
	0.8	236.1	254.6	200.9	207.6	177.2	178.7	145.1	146.4	118	119.2
0.5	0.2	144.1	157.7	137.5	143.2	135.9	137.3	121.6	123	107.2	108.7
	0.4	199.7	217.4	178.1	185	164.1	165.7	138	139.5	114.7	116.1
	0.6	235.9	255	202.4	209.5	179.2	180.7	146.3	147.7	118.4	119.6
	0.8	239.9	257.3	202.8	209	177.1	178.5	144.8	146	118	119.1
0.6	0.2	152.8	167.4	145.8	152	143.7	145.2	127.1	128.7	110.1	111.7
	0.4	208.3	226.8	185.9	193.1	170.9	172.5	142.6	144.2	117	118.4
	0.6	244.4	263.9	209.3	216.5	184.4	185.9	149.6	151	120	121.2
	0.8	247.4	264.5	207.8	213.9	179.9	181.2	146.4	147.5	118.8	119.8
0.7	0.2	162.4	179	155.4	162.5	153.5	155.3	134.2	136	113.6	115.4
	0.4	219.4	239.8	196.4	204.5	180.8	182.7	149.4	151.2	120.2	121.8
	0.6	256.9	278.2	220.2	228.1	193.5	195.2	155.5	157	122.7	124
	0.8	260.2	278.1	217.4	223.7	186.6	187.9	150.3	151.4	120.6	121.6
0.8	0.2	174	193.7	167.7	176.3	167.6	170	144.8	147.2	118.8	121.1
	0.4	234.4	258.8	211.6	221.5	197	199.4	160.8	163.1	125.6	127.7
	0.6	276.2	301.5	238.1	247.6	210.3	212.4	166.4	168.4	127.7	129.3
	0.8	282.7	303.4	235.4	242.7	200.7	202.2	158.8	160.2	124.3	125.5
0.9	0.2	188.9	214.8	185.3	197.3	192.5	196.4	164.9	169	129.5	133.5
	0.4	256.3	289.3	236.3	250.5	228.7	232.8	184.9	188.9	137.7	141.3
	0.6	308.2	343.8	271.1	285.3	247	250.7	192.3	195.6	140	142.8
	0.8	327.2	357.4	275.1	286.1	236.6	239	181.7	183.8	134.7	136.5

** does not satisfy condition (4.5.18).

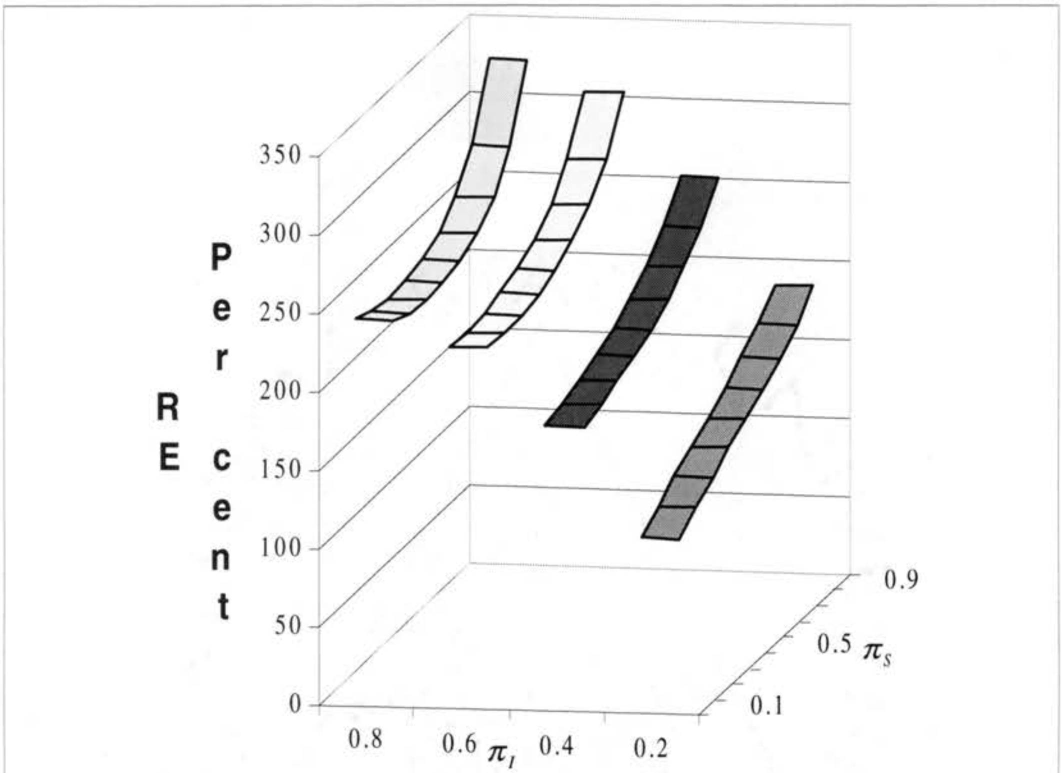

Figure 4.3. The Percent Relative Efficiency of $Var(\hat{\pi}_{UM})/Var(\hat{\pi}_{mS})$

When $P_1 = Q_1 = 0.5$ and $Q_2 = 0.8$.

In Chapter III, it was showed that the variance of an estimator in a stratified Warner's RR model decreases as the number of strata increases. We can apply it to our stratified mixed RR model. Suppose that k strata of equal size are created such that $w_i = 1/k$. Inserting $w_i = 1/k$ into equation (4.5.17), then

$$Var(\hat{\pi}_{mS}) = \frac{1}{n}\left[\sum_{h=1}^{k} w_h \sqrt{L(\hat{\pi}_{S_h}, Q_h, \lambda_h)}\right]^2 = \frac{1}{nk^2}\left[\sum_{h=1}^{k} \sqrt{L(\hat{\pi}_{S_h}, Q_h, \lambda_h)}\right]^2$$

where $L(\hat{\pi}_{S_h}, Q_h, \lambda_h) = \pi_{S_h}(1-\pi_{S_h}) + \dfrac{(1-Q_h)\{\lambda_h Q_h (1-\pi_{S_h}) + (1-\lambda_h)\}}{Q_h^{\,2}}$.

Let $f(k) = \frac{1}{k^2}\left[\sum_{h=1}^{k}\sqrt{L(\hat{\pi}_{S_h}, Q_h, \lambda_h)}\right]^2$ where k is a positive interger.

We want to show that $f(k) - f(k+1) \geq 0$. Then

$$f(k) - f(k+1) = \frac{1}{k^2}\left[\sum_{h=1}^{k}\sqrt{L(\hat{\pi}_{S_h}, Q_h, \lambda_h)}\right]^2 - \frac{1}{(k+1)^2}\left[\sum_{h=1}^{k+1}\sqrt{L(\hat{\pi}_{S_h}, Q_h, \lambda_h)}\right]^2$$

$$= \left[\frac{1}{k}\sum_{h=1}^{k}\sqrt{L(\hat{\pi}_{S_h}, Q_h, \lambda_h)} + \frac{1}{k+1}\sum_{h=1}^{k+1}\sqrt{L(\hat{\pi}_{S_h}, Q_h, \lambda_h)}\right]$$

$$\times\left[\frac{1}{k}\sum_{h=1}^{k}\sqrt{L(\hat{\pi}_{S_h}, Q_h, \lambda_h)} - \frac{1}{k+1}\sum_{h=1}^{k+1}\sqrt{L(\hat{\pi}_{S_h}, Q_h, \lambda_h)}\right].$$

As the number of strata increases, it may be possible to divide a heterogeneous population into subpopulations, each of which is more homogeneous. So we may get

$$\left[\frac{1}{k}\sum_{h=1}^{k}\sqrt{L(\hat{\pi}_{S_h}, Q_h, \lambda_h)} - \frac{1}{k+1}\sum_{h=1}^{k+1}\sqrt{L(\hat{\pi}_{S_h}, Q_h, \lambda_h)}\right] \geq 0$$

By this assumption, $f(k)$ is a monotone decreasing function of k. Thus the variance of an estimator decreases as the number of strata increases. For a case with more than two strata, we may get the same result in terms of efficiency as those for two strata.

4.6. Discussion

A privacy problem of Moors (1971) model discussed by Mangat et al. (1997) and Singh et al. (2000) motivated the authors to present a mixed randomized response model. We showed that the proposed model could rectify the privacy problem of the Moors' model. Furthermore, the mixed model can be more efficient than the Moors model if the "Yes" proportion from the innocuous trait direct is more than half percent. We extended the mixed model to a stratified mixed RR model. We showed that a stratified mixed RR model is more efficient than a mixed RR model in the case with two strata in the population. We derived condition (4.5.18) which makes the stratified mixed RR model more efficient than the Moors model. We conclude that our mixed RR model and stratified mixed RR model are a good alternative models to the Moors model while keeping respondents' confidentiality.

CHAPTER V

A NEW MULTINOMIAL DISTRIBUTION APPROACH TO QUANTITATIVE RANDOMIZED RESPONSE MODEL

5.1. Introduction

Since the introduction of the randomized response technique by Warner (1965), the theory and technique for randomized response (RR) technique have been considerably developed. Abul-Ela et al.(1967) extended Warner's dichotomous RR technique to a polychotomous RR technique but the Abul-Ela et al. RR technique had a drawback. The drawback is that the complexity of the estimation procedure increases as the number of categories in the polychotomy increases. There has been much research on enhancing RR techniques for polichotomies. In particular, Greenberg et al. (1971) adapted the unrelated question qualitative RR technique of Horvitz et al. (1967) to produce the unrelated question quantitative RR technique. A number of quantitative RR techniques have been proposed since Greenberg's quantitative RR technique.

Bourke and Dalenius (1976) presented some new ideas in the realm of randomized response. They pointed out that Greenberg's quantitative RR technique leads to the loss of useful information on the sensitive trait because of the unrelated or nonsensitive question in the quantitative RR technique. To deal with the disadvantage of Greenberg's quantitative RR technique, Eriksson (1973) and Liu and Chow (1976) presented discrete quantitative RR techniques which modified the Greenberg quantitative RR technique. Kim and Flueck (1978) and Himmelfarb and Edgell (1980) developed the additive model approach to RR technique. Pollock and Bek (1976) and Eichhorn and Hayre (1983) introduced the multiplicative RR technique which is the method where a respondent multiplies his or her answer to the sensitive question by a random number from a known distribution. Therefore a validation check for RR technique has also been attempted by Abernathy et al. (1970), Bradburn and Sudman (1979), Tracy and Fox (1981), Danermark and Swensson (1987), Duffy and Waterton (1988) and Kerkvliet (1994). These researchers compared RR interviews and direct interviews based on a statistical measure of efficiency and respondents' protection.

5.2. Proposed Model

5.2.1. The Estimation of Proportions in a Multinomial Distribution

Our RR technique utilizes the Hopkins' device to estimate a multinomial distribution for a sensitive variable (A). Thus our new quantitative RR technique follows the same procedure as Liu and Chow's (1976) RR technique. There are two different colors of balls, red and green, in the device. Each of the green balls has a discrete number marked on it, $0,1,2,\cdots,k+1$. Suppose that all green balls consist of a set of non-sensitive categories, $B=\{B_1,B_2,\cdots,B_{k+1}\}$, such that all the values of A are included.

With t different interviewees performing the Hopkins' device, each interviewee belongs to one of $k+1$ mutually exclusive and exhaustive categories $T=\{T_1,T_2,\cdots,T_{k+1}\}$ which consist of sensitive categories $A=\{A_1,A_2,\cdots,A_{k+1}\}$ and non-sensitive categories $B=\{B_1,B_2,\cdots,B_{k+1}\}$. Let t_i denote the number of observations in a category T_i so that $t=\sum_{i=1}^{k+1} t_i$. We let a_i be the number of observations in a category A_i so that $a=\sum_{i=1}^{k+1} a_i$ and b_i be the number of observations in a category B_i so that

$b = \sum_{i=1}^{k+1} b_i$. We assume that $T_i = t_i$ is the sum of $A_i = a_i$ and $B_i = b_i$. Thus we are attempting to estimate $P_{a1}, P_{a2}, \ldots, P_{a(k+1)}$ the proportions in the population who are in sensitive categories $A_1, A_2, \cdots, A_{k+1}$. Based on green balls with number in the Hopkins' device, we can derive the proportions in the population who are in categories $B_1, B_2, \cdots, B_{k+1}$ by $P_{bi} = g_i / g$.

Let $P_{t1}, P_{t2}, \ldots, P_{t(k+1)}$ denote the proportions in the population who are in categories $T_1, T_2, \cdots, T_{k+1}$. When t different interviewees finish performing the Hopkins' device, we can derive b the total number of people who are in $B = \{B_1, B_2, \cdots, B_{k+1}\}$ by $b \geq tg/(r+g)$ where b is an integer. The condition $b \geq tg/(r+g)$ does not influence the result of the column B in Table 5.1 when t is a large enough number.

We can also derive $b_1, b_2, \ldots, b_k$ in the same way that $b_i \geq tg_i/(r+g_i)$ where b_i is an integer. Thus, we can get $b_{k+1} = b - (b_1 + b_2 + \ldots + b_k)$.

TABLE 5.1.

The Number of Observations for Three Different Variables.

	T	A	B
Category 1	$T_1 = t_1$	$A_1 = a_1$	$B_1 = b_1$
Category 2	$T_2 = t_2$	$A_2 = a_2$	$B_2 = b_2$
Category 3	$T_3 = t_3$	$A_3 = a_3$	$B_3 = b_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Category k	$T_k = t_k$	$A_k = a_k$	$B_k = b_k$
Category k+1	$T_{k+1} = t_{k+1}$	$A_{k+1} = a_{k+1}$	$B_{k+1} = b_{k+1}$
Total	t	a	b

Then we can define a multinomial distribution of T, A and B as follow:

$$T = (T_1, T_2, \ldots, T_k) \sim \text{MULT}(t, P_{t1,} P_{t2}, \ldots, P_{tk})$$

$$A = (A_1, A_2, \ldots, A_k) \sim \text{MULT}(a, P_{a1,} P_{a2}, \ldots, P_{ak}) \qquad (5.2.1)$$

$$B = (B_1, B_2, \ldots, B_k) \sim \text{MULT}(b, P_{b1,} P_{b2}, \ldots, P_{bk}).$$

Suppose that $T = A + B$ and respondents give truthful answers to one of two different questions. From the moment generating functions of T, A and B or directly from the marginal probability mass function's, we can compute moments.

$$E(T_h) = tP_{th}, \; E(A_h) = aP_{ah} \text{ and } E(B_h) = bP_{bh} \qquad \text{where } h = 1, 2, \ldots, k+1. \qquad (5.2.2)$$

For $T_h = A_h + B_h$,

$$E(A_h) = E(T_h) - E(B_h) = tP_{th} - bP_{bh}. \qquad (5.2.3)$$

Since $E(A_h) = aP_{ah}$,

$$P_{ah} = \frac{tP_{th} - bP_{bh}}{a} = \frac{tP_{th} - bP_{bh}}{t - b}. \qquad (5.2.4)$$

Let $\hat{P}_{ah}$ denote the estimate of P_{ah} and $\hat{P}_{th}$ denote the estimate of P_{th}. Since $\hat{P}_{bh} = g_h / g$,

$$\hat{P}_{ah} = \frac{t\hat{P}_{th} - b(g_h / g)}{t - b} \qquad (5.2.5)$$

which is an unbiased estimator of P_{ah}. The estimate of variance is

$$v(\hat{P}_{ah}) = \frac{t\hat{P}_{th}(1 - \hat{P}_{th})}{(t - b)^2}. \qquad (5.2.6)$$

The estimate of covariance is

$$C\hat{o}v(\hat{P}_{ah}, \hat{P}_{ai}) = -\frac{t\hat{P}_{th}\hat{P}_{ti}}{(t - b)^2} \qquad \text{where } h \neq i. \qquad (5.2.7)$$

5.2.2. A Random Transformation to the True Estimate

In the previous section, we assumed that respondents report truthfully. But in a case of untruthful reporting, we need to derive an estimator for population proportion P_{ah} with the prior information from (5.2.5), (5.2.6) and (5.2.7) when a respondent reports untruthfully. Let R_{ij} denote the probability that a person of category i announces himself or herself as one of category j. Suppose that respondents report truthfully when they have a non-sensitive question. Then we can apply the lying model of Mukhopadhyay (1980) to the sensitive question. Assume that there is a sensitive category A_i for $i = 1,2,3,4$ such that A_1 has no social stigma and that there is more social stigma as i increases. Intuitively, we can stipulate the following:

$$R_{12} = R_{13} = R_{14} = R_{23} = R_{24} = R_{34} = 0,\ \ R_{11} = 1, R_{21} + R_{22} = 1,$$

$$R_{31} + R_{32} + R_{33} = 1 \ \text{ and } \ R_{41} + R_{42} + R_{43} + R_{44} = 1. \tag{5.2.8}$$

Let π_i represent the true proportion of respondents who belong to a sensitive category i and P_{ai} represent the observed proportion of respondents who belong to a sensitive category i. Under the these assumptions, we can derive the following:

$$\begin{aligned}
P_{a1} &= R_{11}\pi_1 + R_{21}\pi_2 + R_{31}\pi_3 + R_{41}\pi_4 = \pi_1 + R_{21}\pi_2 + R_{31}\pi_3 + R_{41}\pi_4 \\
P_{a2} &= R_{12}\pi_1 + R_{22}\pi_2 + R_{32}\pi_3 + R_{42}\pi_4 = R_{22}\pi_2 + R_{32}\pi_3 + R_{42}\pi_4 \\
P_{a3} &= R_{13}\pi_1 + R_{23}\pi_2 + R_{33}\pi_3 + R_{43}\pi_4 = R_{33}\pi_3 + R_{43}\pi_4 \\
P_{a4} &= R_{14}\pi_1 + R_{24}\pi_2 + R_{34}\pi_3 + R_{44}\pi_4 = R_{44}\pi_4.
\end{aligned} \tag{5.2.9}$$

Then

$$P=\begin{pmatrix}P_{a1}\\P_{a2}\\P_{a3}\\P_{a4}\end{pmatrix}=\begin{pmatrix}1 & R_{21} & R_{31} & R_{41}\\0 & R_{22} & R_{32} & R_{42}\\0 & 0 & R_{33} & R_{43}\\0 & 0 & 0 & R_{44}\end{pmatrix}\begin{pmatrix}\pi_1\\\pi_2\\\pi_3\\\pi_4\end{pmatrix}$$

$$=\begin{pmatrix}1 & R_{21} & R_{31} & R_{41}\\0 & 1-R_{21} & R_{32} & R_{42}\\0 & 0 & 1-R_{31}-R_{32} & R_{43}\\0 & 0 & 0 & 1-R_{41}-R_{42}-R_{43}\end{pmatrix}\begin{pmatrix}\pi_1\\\pi_2\\\pi_3\\\pi_4\end{pmatrix}. \tag{5.2.10}$$

We can extend the four-category sensitive case to the k -category sensitive case. Assume that there is a sensitive category A_i for $i=1,2,\ldots,k$ such that A_1 has no social stigma and A_i is more social stigma as i increases. Like (5.2.8), we can stipulate the following:

$$R_{ij}=0 \quad \text{if} \quad i<j \quad \text{where } i,j=1,2,\ldots,k$$

$$\sum_{j=1}^{k} R_{ij}=1 \quad \text{for } i=1,2,\ldots,k. \tag{5.2.11}$$

Like (5.2.9), we can derive the following:

$$P_{aj}=\sum_{i=1}^{k} R_{ij}\pi_i \quad \text{for } j=1,2,\ldots,k. \tag{5.2.12}$$

Then

$$P=\begin{pmatrix}P_{a1}\\P_{a2}\\\vdots\\P_{a(k-1)}\\P_{ak}\end{pmatrix}=\begin{pmatrix}1 & R_{21} & \cdots & R_{(k-1)1} & R_{k1}\\0 & 1-R_{21} & \cdots & R_{(k-1)2} & R_{k2}\\\vdots & \vdots & \vdots & \vdots & \vdots\\0 & 0 & 0 & 1-\sum_{j=1}^{k-2}R_{(k-1)j} & R_{k(k-1)}\\0 & 0 & 0 & 0 & 1-\sum_{j=1}^{k-1}R_{kj}\end{pmatrix}\begin{pmatrix}\pi_1\\\pi_2\\\vdots\\\pi_{k-1}\\\pi_k\end{pmatrix} \tag{5.2.13}$$

We can rewrite it like this:

$$P = R\pi$$

where $\pi = (\pi_1 \quad \pi_2 \quad \cdots \quad \pi_{k-1} \quad \pi_k)^T$ and

$$R = \begin{pmatrix} 1 & R_{21} & \cdots & R_{(k-1)1} & R_{k1} \\ 0 & 1-R_{21} & \cdots & R_{(k-1)2} & R_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1-\sum_{j=1}^{k-2} R_{(k-1)j} & R_{k(k-1)} \\ 0 & 0 & 0 & 0 & 1-\sum_{j=1}^{k-1} R_{kj} \end{pmatrix}$$

If R is nonsingular, we can derive the true proportions for sensitive categories:

$$\pi = R^{-1}P .$$

The Maximum Likelihood estimator of $\pi = (\pi_1 \quad \pi_2 \quad \cdots \quad \pi_{k-1} \quad \pi_k)^T$ is given by

$$\hat{\pi} = R^{-1}\hat{P} .$$

where $\hat{P}$ is a estimate vector of P, provided that the vector $\hat{\pi}$ satisfies $\hat{\pi}_i \geq 0$ for all i and $\sum_{i=1}^{4} \hat{\pi}_i = 1$.

The estimate of covariance of $\hat{\pi}$ is

$$C\hat{o}v(\hat{\pi}) = R^{-1}C\hat{o}v(\hat{P})(R^{-1})^T$$

where $$C\hat{o}v(\hat{P}_{ai}, \hat{P}_{aj}) = \begin{cases} \dfrac{t\hat{P}_{ti}(1-\hat{P}_{ti})}{(t-b)^2} & if\ i = j \\ -\dfrac{t\hat{P}_{ti}\hat{P}_{tj}}{(t-b)^2} & if\ i \neq j \end{cases} .$$

5.3. Large Sample Multiple Comparisons for RR Model

We are interested in investigating a multiple contrast method for sensitive category proportions of multinomial populations. Let $\pi_j^{(i)}$ be the true proportion of respondents who belong to a sensitive category j in the ith multinomial population ($i=1,2,\ldots,m$ and $j=1,2,\ldots,k+1$) such that $\sum_{j=1}^{k+1}\pi_j^{(i)}=1$ for all i. From Goodman (1964), defining a contrast to be $\varphi=\sum_{i\,j}\alpha_{ij}\pi_j^{(i)}$ where $\sum_{i=1}^{m}\alpha_{ij}=0$ for all j. Let $a^{(i)}$ denote the total number of observations in all sensitive categories in the i th multinomial population and $a_j^{(i)}$ denote the number of observations in a sensitive category A_j in the ith multinomial population. Then $a^{(i)}=a_1^{(i)}+a_2^{(i)}+\ldots+a_{k+1}^{(i)}$ for all i. We denote $a_j^{(i)}/a^{(i)}$ by $p_j^{(i)}$. The Maximum Likelihood (ML) estimator of $\pi_j^{(i)}$ is $p_j^{(i)}$. Then the ML estimator of $\varphi=\sum_{i\,j}\alpha_{ij}\pi_j^{(i)}$ is $\hat{\varphi}=\sum_{i\,j}\alpha_{ij}p_j^{(i)}$. The variance of $p_j^{(i)}$ is $\pi_j^{(i)}(1-\pi_j^{(i)})/a^{(i)}$, and the variance of $\hat{\varphi}$ is

$$Var(\hat{\varphi})=\sum_{i=1}^{m}\left\{\left[\sum_{j=1}^{k+1}\alpha_{ij}^2\pi_j^{(i)}-\left(\sum_{j=1}^{k+1}\alpha_{ij}\pi_j^{(i)}\right)^2\right]\Big/a^{(i)}\right\}. \tag{5.3.1}$$

These two variances can be estimated by $p_j^{(i)}(1-p_j^{(i)})/a^{(i)}$ and

$$v(\hat{\varphi})=\sum_{i=1}^{m}\left\{\left[\sum_{j=1}^{k+1}\alpha_{ij}^2 p_j^{(i)}-\left(\sum_{j=1}^{k+1}\alpha_{ij}p_j^{(i)}\right)^2\right]\Big/a^{(i)}\right\}. \tag{5.3.2}$$

When the α_{ij} have been specified a priori and $a^{(i)}\to\infty$, the probability will approach $1-\alpha$ that

$$\hat{\varphi}-Z_{\alpha/2}\sqrt{v(\hat{\varphi})}\le\varphi\le\hat{\varphi}+Z_{\alpha/2}\sqrt{v(\hat{\varphi})}\,, \tag{5.3.3}$$

here $Z_{\alpha/2}$ is the $100(1-\alpha)$ th percentile of the unit normal distribution.

If the Pearson χ^2 test statistics leads to failure of the rejection of the null hypothesis $H_0: \pi_j^{(1)} = \pi_j^{(2)} = \cdots = \pi_j^{(m)} = \pi_j$ which means that m multinomial populations are homogeneous, all simultaneous confidence intervals would include zero. But if the test leads to rejection of this null hypothesis, we can use the multiple contrast method of Goodman (1964) to determine which particular contrasts are significantly different from zero. First, we need the test of homogeneity using the Pearson χ^2 test statistics for m-sample multinomal.

Example 1. Suppose that we apply our new RR technique to three different female sample groups to research abortion and we obtain the three sensitive categories from three female groups. We assume that the result of the experiment is Table 5.3. First, we want to test the null hypothesis $H_0: \pi_j^{(1)} = \pi_j^{(2)} = \pi_j^{(3)} = \pi_j$. We used the Pearson χ^2 test statistics for 3 multinomial samples. The test statistics is

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{(a_j^{(i)} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = 11.88 > 9.49 = \chi^2_{.95}(4). \tag{5.3.4}$$

TABLE 5.2

The Proportions of Respondents Who Belong to Each Sensitive Category.

	A_1	A_2	A_3	
Group1	$\pi_1^{(1)}$	$\pi_2^{(1)}$	$\pi_3^{(1)}$	1
Group2	$\pi_1^{(2)}$	$\pi_2^{(2)}$	$\pi_3^{(2)}$	1
Group3	$\pi_1^{(3)}$	$\pi_2^{(3)}$	$\pi_3^{(3)}$	1
	π_1	π_2	π_3	1

TABLE 5.3

Observed and Estimated Expected Outcomes for a Three-sample.

	Observed outcomes (Estimated Expected Outcomes)			
	A_1	A_2	A_3	Total
Female Group1	55(57)	10(10.8)	10(7.2)	75
Female Group2	75(76)	21(14.4)	4(9.6)	100
Female Group3	60(57)	5(10.8)	10(7.2)	75
Total	190	36	24	250

H_0 can be rejected at $\alpha = 0.05$ level. Thus we want to know which particular contrasts are significantly different from zero. Applying the LSD multiple contrast method to three female group proportions, then the 95% set of confidence intervals about two different proportions is given by

$$\left(p_j^{(h)} - p_j^{(i)}\right) - Z_{0.025}\sqrt{p_j^{(h)}(1-p_j^{(h)})/a^{(h)} + p_j^{(i)}(1-p_j^{(i)})/a^{(l)}} \; < \; \pi_j^{(h)} - \pi_j^{(i)}$$

$$< \left(p_j^{(h)} - p_j^{(i)}\right) + Z_{0.025}\sqrt{p_j^{(h)}(1-p_j^{(h)})/a^{(h)} + p_j^{(i)}(1-p_j^{(i)})/a^{(l)}} \qquad (5.3.5)$$

where $h \neq i$ and $p_j^{(i)}$ is the MLE of $\pi_j^{(i)}$. The following nine contrast comparisons are as follows:

$-0.1567 < \pi_1^{(1)} - \pi_1^{(2)} < 0.1234$ Not Significant

$-0.2016 < \pi_1^{(1)} - \pi_1^{(3)} < 0.0683$ Not Significant

$-0.1834 < \pi_1^{(2)} - \pi_1^{(3)} < 0.0834$ Not Significant

$-0.1806 < \pi_2^{(1)} - \pi_2^{(2)} < 0.0273$ Not Significant

$-0.016 < \pi_2^{(1)} - \pi_2^{(3)} < 0.1493$ Not Significant

$0.0497 < \pi_2^{(2)} - \pi_2^{(3)} < 0.2369$		Significant
$0.0045 < \pi_3^{(1)} - \pi_3^{(2)} < 0.1821$		Significant
$-0.1088 < \pi_3^{(1)} - \pi_3^{(3)} < 0.1088$		Not Significant
$-0.1821 < \pi_3^{(2)} - \pi_3^{(3)} < -0.0045$		Significant

From the above information, these three pairwise contrasts $\pi_2^{(3)} < \pi_2^{(2)}$, $\pi_3^{(2)} < \pi_3^{(1)}$ and $\pi_3^{(2)} < \pi_3^{(3)}$ are significantly different. Except for the three pairwise contrasts, the other six contrasts are not significantly different from zero.

5.4. Correlation between Two Different Sensitive Questions

Fox and Tracy (1984) considered estimating the correlation between two sensitive variables which are surveyed under the quantitative RR technique by Greenberg et al. (1971). In this paper, we will consider estimating the correlation between two sensitive variables which is based on a new quantitative RR technique. For an interview involving two sensitive questions, a researcher prepares two Hopkins' devices which have different ratios of red balls and green balls with designated numbers. An interviewee will face two devices so that she or he will use different device for each sensitive question independently. For each question, the respondent will shake the device and will get a ball. If the ball is red then the respondent should answer the sensitive question. Otherwise, if the ball is green with a designated number then the respondent will just say the number on the green ball. For two different questions, we are going to use the multivariate randomized response design of Bourke (1981). We denote θ_{ij} to be the probability that a respondent gives the ith category for the first question and the j th category for the second question. Let P_{ij} denote the true proportion of respondents who fall in the i th category for the first question and the j th category for the second question. Suppose that the first question has I categories and the second question has J categories. For the conditional probability $P[k\,l \mid i\,j]$ that a respondent of category i and category j announces himself or herself as one of category k and category l, we have

$$\theta_{ij} = \sum_{i=1}^{I}\sum_{j=1}^{J} P[k\,l \mid i\,j]\lambda_{ij} \tag{5.4.1}$$

where λ_{ij} is the true proportion that a respondent belongs in the i th category for the first question and belongs in the j th category for the second question. Since two devices are independently performed by a respondent, we can write $P[k\,l \mid i\,j] = P_1[k \mid i]P_2[l \mid j]$. Therefore (5.4.1) can be rewritten like this:

$$\theta_{ij} = \sum_{i=1}^{I}\sum_{j=1}^{J} P_1[k \mid i]P_2[l \mid j]\lambda_{ij}\,. \tag{5.4.2}$$

By Bourke (1981), we can express the vectors $\theta^{(2)}$ and $\lambda^{(2)}$ so that θ_{ij} is the r th element of the vector $\theta^{(2)}$ and λ_{ij} is the c th element of the vector $\lambda^{(2)}$, where

$$r = l + (k-1)I\,,\ \ c = j + (i-1)J\,. \tag{5.4.3}$$

We can express a matrix $M^{(2)}$ so that $P[k\,l \mid i\,j] = P[k \mid i]P[l \mid j]$ is in the (r,c) position of the matrix $M^{(2)}$. The $M^{(2)}$ is the Kronecker product of two matrixes M_1 and M_2 so that $P[k \mid i]$ is a element of M_1 and $P[l \mid j]$ is a element of M_2. Therefore the vector $\theta^{(2)}$ can be expressed as follows:

$$\theta^{(2)} = M^{(2)}\lambda^{(2)} = (M_1 \otimes M_2)\lambda^{(2)}\,. \tag{5.4.4}$$

If $M_1 \otimes M_2$ is nonsingular, we can derive $\lambda^{(2)}$ from (5.4.4) as follows:

$$\lambda^{(2)} = (M_1 \otimes M_2)^{-1}\theta^{(2)}. \tag{5.4.5}$$

If $\hat{\theta}^{(2)}$ is the asymptotic Maximum Likelihood estimator of $\theta^{(2)}$ then we can estimate

$$\hat{\lambda}^{(2)} = (M_1 \otimes M_2)^{-1}\hat{\theta}^{(2)}. \tag{5.4.6}$$

Using these cell proportions $\hat{\lambda}^{(2)}$, we can consider the product moment correlation between two sensitive variables ($A^{(1)}$ and $A^{(2)}$). From the interview, we can directly estimate the Pearson product-moment correlation between two different variables ($T^{(1)} = A^{(1)} + B^{(1)}$ and $T^{(2)} = A^{(2)} + B^{(2)}$). When $T^{(1)}$ is a row variable and $T^{(2)}$ is a column variable, we let $A(r_i)$ denote a value assigned to the ith row category, and $A(c_j)$ denote a value assigned to the jth column category.

Suppose $A(r_1) \le A(r_2) \le \cdots \le A(r_I)$ and $A(c_1) \le A(c_2) \le \cdots \le A(c_J)$.

For $I \times J$ contingency table

$$\rho_T = \frac{Cov\left(T^{(1)}, T^{(2)}\right)}{\sqrt{Var(T^{(1)})}\sqrt{Var(T^{(2)})}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\lambda_{ij}\left(A(r_i) - A(\bar{r})\right)\left(A(c_j) - A(\bar{c})\right)}{\sqrt{\left\{\sum_{i=1}^{I}\lambda_{i+}\left(A(r_i) - A(\bar{r})\right)^2\right\}\left\{\sum_{j=1}^{J}\lambda_{+j}\left(A(c_j) - A(\bar{c})\right)^2\right\}}}$$

(5.4.7)

where $\lambda_{i+} = \sum_{j=1}^{J}\lambda_{ij}$, and $\lambda_{+j} = \sum_{i=1}^{I}\lambda_{ij}$, and $A(\bar{r}) = \sum_{i=1}^{I}\lambda_{i+}A(r_i)$ and $A(\bar{c}) = \sum_{j=1}^{J}\lambda_{+j}A(c_j)$.

The estimator is

$$r_T = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{\lambda}_{ij}\left(A(r_i)-A(\hat{\bar{r}})\right)\left(A(c_j)-A(\hat{\bar{c}})\right)}{\sqrt{\left\{\sum_{i=1}^{I}\hat{\lambda}_{i+}\left(A(r_i)-A(\hat{\bar{r}})\right)^2\right\}\left\{\sum_{j=1}^{J}\hat{\lambda}_{+j}\left(A(c_j)-A(\hat{\bar{c}})\right)^2\right\}}} \tag{5.4.8}$$

where $A(\hat{\bar{r}}) = \sum_{i=1}^{I}\left(\hat{\lambda}_{i+}/\hat{\lambda}_{++}\right)A(r_i)$ and $A(\hat{\bar{c}}) = \sum_{j=1}^{J}\left(\hat{\lambda}_{+j}/\hat{\lambda}_{++}\right)A(c_j)$.

Since $A^{(1)}$ and $B^{(1)}$ are independent, and $A^{(2)}$ and $B^{(2)}$ are independent. Then

$Var(T^{(1)}) = Var(A^{(1)}) + Var(B^{(1)})$ and $Var(T^{(2)}) = Var(A^{(2)}) + Var(B^{(2)})$.

Suppose $A^{(1)}$ and $B^{(2)}$, $A^{(2)}$ and $B^{(1)}$, and $B^{(1)}$ and $B^{(2)}$ are uncorrelated each other.

Then the covariance of two variable $T^{(1)}$ and $T^{(2)}$ is

$$\begin{aligned} Cov(T^{(1)}, T^{(2)}) &= Cov(A^{(1)} + B^{(1)}, A^{(2)} + B^{(2)}) \\ &= Cov(A^{(1)}, A^{(2)}) + Cov(A^{(1)}, B^{(2)}) + Cov(B^{(1)}, A^{(2)}) + Cov(B^{(1)}, B^{(2)}) \\ &= Cov(A^{(1)}, A^{(2)}). \end{aligned}$$

The product moment correlation between two sensitive variables $A^{(1)}$ and $A^{(2)}$ is

$$\rho_A = \frac{Cov\left(A^{(1)}, A^{(2)}\right)}{\sqrt{Var(A^{(1)})}\sqrt{Var(A^{(2)})}} = \frac{Cov\left(T^{(1)}, T^{(2)}\right)}{\sqrt{Var(T^{(1)}) - Var(B^{(1)})}\sqrt{Var(T^{(1)}) - Var(B^{(1)})}}$$

$$= \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\lambda_{ij}\left(A(r_i)-A(\bar{r})\right)\left(A(c_j)-A(\bar{c})\right)}{\sqrt{\left\{\sum_{i=1}^{I}\lambda_{i+}\left(A(r_i)-A(\bar{r})\right)^2 - Var\left(B^{(1)}\right)\right\}\left\{\sum_{j=1}^{J}\lambda_{+j}\left(A(c_j)-A(\bar{c})\right)^2 - Var\left(B^{(2)}\right)\right\}}}. \tag{5.4.9}$$

From (2.2.6), (2.2.7) and (5.2.1), we can derive the mean and variance of a designated number i:

$$\mu_B = \sum_{i=1}^{k} i \frac{g_i}{1-P} \quad \text{and} \quad Var(B) = \sum_{i=1}^{k} (i-\mu_B)^2 \frac{g_i}{1-P} \tag{5.4.10}$$

where the proportion of green balls with designated number i is g_i such that $1-P=\sum_{i=1}^{m} g_i$. The estimator of ρ_A is

$$r_A = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{\lambda}_{ij}\left(A(r_i)-A(\hat{\bar{r}})\right)\left(A(c_j)-A(\hat{\bar{c}})\right)}{\sqrt{\left\{\sum_{i=1}^{I}\hat{\lambda}_{i+}\left(A(r_i)-A(\hat{\bar{r}})\right)^2 - Var\left(B^{(1)}\right)\right\}\left\{\sum_{j=1}^{J}\hat{\lambda}_{+j}\left(A(c_j)-A(\hat{\bar{c}})\right)^2 - Var\left(B^{(2)}\right)\right\}}} \tag{5.4.11}$$

where $A(\hat{\bar{r}})=\sum_{i=1}^{I}\left(\hat{\lambda}_{i+}/\hat{\lambda}_{++}\right)A(r_i)$ and $A(\hat{\bar{c}})=\sum_{j=1}^{J}\left(\hat{\lambda}_{+j}/\hat{\lambda}_{++}\right)A(c_j)$.

If the value of r_A equals zero then it means that two sensitive variables $A^{(1)}$ and $A^{(2)}$ are independent. The farther the absolute value of r_A is from zero, the stronger the relationship between two sensitive variables $A^{(1)}$ and $A^{(2)}$ correlate with each other.

TABLE 5.4

The Number of Respondents Who Belongs to Two Different Sensitive Categories.

	Observed Outcomes			
	A(c₁)=1	***A(c₂)=2***	***A(c₃)=3***	***Total***
A(r₁)=1	***45***	***9***	***6***	***60***
A(r₂)=2	***18***	***5***	***2***	***25***
A(r₃)=3	***10***	***3***	***2***	***15***
	73	***17***	***10***	***100***

Example 2. Suppose that we use our new RR technique to know the correlation between two different sensitive variables and we set the three sensitive categories from each of two sensitive variables. We assume that the result of the experiment is Table 5.4. From (5.4.1) and (5.4.2), the proportion that a respondent of category i and category j announces himself or herself as one of category k and one of category l is

$$\theta_{ij} = \sum_{i=1}^{3}\sum_{j=1}^{3} P[k\,l \mid i\,j]\lambda_{ij} = \sum_{i=1}^{3}\sum_{j=1}^{3} P_1[k \mid i]P_2[l \mid j]\lambda_{ij}\,. \tag{5.4.12}$$

From (5.4.4), the probability θ_{ij} can be expressed as follows:

$$\theta^{(2)} = \begin{pmatrix} \theta_{11} \\ \theta_{12} \\ \theta_{13} \\ \theta_{21} \\ \theta_{22} \\ \theta_{23} \\ \theta_{31} \\ \theta_{32} \\ \theta_{33} \end{pmatrix} = M^{(2)}\lambda^{(2)} = \begin{pmatrix} (P_1[1|1]\;P_1[1|2]\;P_1[1|3]) \otimes (P_2[1|1]\;P_2[1|2]\;P_2[1|3]) \\ (P_1[1|1]\;P_1[1|2]\;P_1[1|3]) \otimes (P_2[2|1]\;P_2[2|2]\;P_2[2|3]) \\ (P_1[1|1]\;P_1[1|2]\;P_1[1|3]) \otimes (P_2[3|1]\;P_2[3|2]\;P_2[3|3]) \\ (P_1[2|1]\;P_1[2|2]\;P_1[2|3]) \otimes (P_2[1|1]\;P_2[1|2]\;P_2[1|3]) \\ (P_1[2|1]\;P_1[2|2]\;P_1[2|3]) \otimes (P_2[2|1]\;P_2[2|2]\;P_2[2|3]) \\ (P_1[2|1]\;P_1[2|2]\;P_1[2|3]) \otimes (P_2[3|1]\;P_2[3|2]\;P_2[3|3]) \\ (P_1[3|1]\;P_1[3|2]\;P_1[3|3]) \otimes (P_2[1|1]\;P_2[1|2]\;P_2[1|3]) \\ (P_1[3|1]\;P_1[3|2]\;P_1[3|3]) \otimes (P_2[2|1]\;P_2[2|2]\;P_2[2|3]) \\ (P_1[3|1]\;P_1[3|2]\;P_1[3|3]) \otimes (P_2[3|1]\;P_2[3|2]\;P_2[3|3]) \end{pmatrix} \begin{pmatrix} \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{21} \\ \lambda_{22} \\ \lambda_{23} \\ \lambda_{31} \\ \lambda_{32} \\ \lambda_{33} \end{pmatrix} \tag{5.4.13}$$

By rewriting $M^{(2)}$ in terms of M_1 and M_2, we get

$$M^{(2)} = M_1 \otimes M_2 = \left(\begin{bmatrix} P_1[1|1] & P_1[1|2] & P_1[1|3] \\ P_1[2|1] & P_1[2|2] & P_1[2|3] \\ P_1[3|1] & P_1[3|2] & P_1[3|3] \end{bmatrix} \otimes \begin{bmatrix} P_2[1|1] & P_2[1|2] & P_2[1|3] \\ P_2[2|1] & P_2[2|2] & P_2[2|3] \\ P_2[3|1] & P_2[3|2] & P_2[3|3] \end{bmatrix} \right) \tag{5.4.14}$$

From Table 5.4, we can derive each θ_{ij} as follows:

$$\theta^{(2)} = (.45 \quad .09 \quad .06 \quad .18 \quad .05 \quad .02 \quad .1 \quad .03 \quad .02)^T.$$

Suppose that

$$M_1 = \begin{bmatrix} 1 & .1 & .1 \\ 0 & .9 & .1 \\ 0 & 0 & .8 \end{bmatrix} \text{ and } M_2 = \begin{bmatrix} 1 & .2 & .2 \\ 0 & .8 & .1 \\ 0 & 0 & .7 \end{bmatrix}.$$

$$M = \begin{bmatrix} 1 & .1 & .1 \\ 0 & .9 & .1 \\ 0 & 0 & .8 \end{bmatrix} \otimes \begin{bmatrix} 1 & .2 & .2 \\ 0 & .8 & .1 \\ 0 & 0 & .7 \end{bmatrix} = \begin{pmatrix} 1 & .2 & .2 & .1 & .02 & .02 & .1 & .02 & .02 \\ 0 & .8 & .1 & 0 & .08 & .01 & 0 & .08 & .01 \\ 0 & 0 & .7 & 0 & 0 & .07 & 0 & 0 & .07 \\ 0 & 0 & 0 & .9 & .18 & .18 & .1 & .02 & .02 \\ 0 & 0 & 0 & 0 & .72 & .09 & 0 & .08 & .01 \\ 0 & 0 & 0 & 0 & 0 & .63 & 0 & 0 & .07 \\ 0 & 0 & 0 & 0 & 0 & 0 & .8 & .16 & .16 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .64 & .08 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .56 \end{pmatrix}.$$

Since M is nonsingular and θ_{ij} is known, we can derive the true proportion λ_{ij} that a respondent belongs in the ith category for the first question and belongs in the jth category for the second question. By finding the inverse matrix of M, we get

$$\lambda^{(2)} = M^{-1}\theta^{(2)} = (.385 \quad .091 \quad .079 \quad .168 \quad .06 \quad .028 \quad .109 \quad .042 \quad .036)^T.$$

Suppose that two different sensitive questions are independent each other and the first sensitive variable ($A^{(1)}$) and the first non-sensitive variable ($B^{(2)}$), the second sensitive variable ($A^{(2)}$) and the first non-sensitive variable ($B^{(1)}$), and the first non-sensitive variable ($B^{(1)}$) and the second non-sensitive variable ($B^{(2)}$) are uncorrelated each other. From (5.4.9), the product moment correlation between two sensitive variables $A^{(1)}$ and $A^{(2)}$ is

$$\rho_A = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3}\lambda_{ij}\left(A(r_i)-A(\bar{r})\right)\left(A(c_j)-A(\bar{c})\right)}{\sqrt{\left\{\sum_{i=1}^{3}\lambda_{i+}\left(A(r_i)-A(\bar{r})\right)^2 - Var\left(B^{(1)}\right)\right\}\left\{\sum_{j=1}^{3}\lambda_{+j}\left(A(c_j)-A(\bar{c})\right)^2 - Var\left(B^{(2)}\right)\right\}}}$$

If a researcher uses two Hopkins' devices which have different ratios of red balls and green balls with a designated number then she or he can derive the variances of a designated number i, that is, $Var(B^{(1)})$ and $Var(B^{(2)})$.

Suppose that we get $Var(B^{(1)}) = .567$ and $Var(B^{(2)}) = .479$ from the randomized response technique. Then we can easily compute the correlation between two sensitive variables. The correlation is

$$\rho_A = \frac{.0404}{\sqrt{\{.6076 - .567\}\{.5353 - .479\}}} = .845.$$

It means that the relationship between two sensitive variables $A^{(1)}$ and $A^{(2)}$ correlated strongly each other. Through Example 2, we discover the important fact that if researchers choose two sensitive issues highly correlated then they may obtain more useful information, for example, like the correlation between abortion and alcohol abuse, in addition to get a reliable data.

5.5. Discussion

A multinomial distribution approach to a new RR technique using a Hopkins' device will be introduced. Eriksson (1973) and Liu and Chow (1976) have presented a quantitative randomized response technique which is modified by Greenberg et al. (1971). But the result of their researches focused on estimating the proportions which are the observed estimates of sensitive category proportions. Furthermore they did not apply their randomized response models to the multivariate randomized response design for a sensitive variable. It is advantageous to treat ordinal data in a quantitative manner by assigning ordered scores to the categories. In a new quantitative RR technique, we derived the true proportion estimates of the sensitive categories based on the observed estimates of sensitive category proportions. We applied a multiple contrast method by Goodman (1964) to a randomized response technique. Through a multiple contrast method for a randomized response technique, we might be able to investigate the sensitive trait in a target group population in detail. A Pearson product-moment correlation between two sensitive variables was presented in this research. Since researchers often deal with categorical data of sensitive issues in a real life, the Pearson product-moment correlation is more appropriate than the correlation between two sensitive variables presented by Fox and Tracy (1984). Through the Pearson product-moment correlation presented in this research, researchers may get more useful information of the relationship between two different sensitive questions in the same interview.

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

6.1. Conclusions

The purpose of this dissertation was to develop a new randomized response technique which is more efficient than the previous randomized response techniques while keeping the respondents' confidentiality, and to investigate the properties of that randomized response technique. Under the assumption of respondents' truthful reporting, it was shown that a stratified Warner randomized response model was more efficient than the Warner (1965) RR model, Mangat and Singh (1990) RR model, and Mangat (1994) RR model under the conditions given in this dissertation. In terms of the untruthful reporting case, we showed that a stratified Warner randomized response model was more efficient than the Warner (1965) RR model, Mangat and Singh (1990) RR model under the conditions given in this dissertation.

Furthermore, a mixed randomized response model was more efficient than Moors (1971) RR model and rectified the privacy problem of the Moors model. The author extended the mixed RR model to a stratified mixed RR model. It was concluded that a stratified mixed RR model was more efficient than the mixed RR model.

The last goal was accomplished in Chapter 5. For obtaining ordinal data in a quantitative manner, a multinomial distribution approach to a new RR technique using a Hopkins' randomizing device was introduced. Using the new quantitative RR model, the author tried to investigate the relationship between two sensitive variables by a way of presenting a multiple contrast method and derive the Pearson product-moment correlation between the two sensitive questions.

6.2. Future Work

Bayesian analyses of randomized response models are given in Winkler and Franklin (1979), Pitz (1980), O'Hagan (1987), Oh(1994), and Unnikrishnan and Kunte (1999). Bayesian methods are attractive in randomized response models because they incorporate useful prior information where only partial information is available. Most of research on the Bayesian approach to randomized response models focuses on dichotomous and polychotomous responses. Research is needed on a Bayesian approach to a quantitative randomized response model since researchers more often deal with sensitive issues of a quantitative character in practical fields. A Bayesian approach to a quantitative randomized response model will be useful and practical.

BIBLIOGRAPHY

Abernathy, J.R., Greenberg, B.G., and Horvitz, D.G. (1970), "Estimates of induced abortion in urban North Carolina," *Demography*, 7, 19-29.

Abul-Ela, A.A., Greenberg, B.G., and Horvitz, D.G. (1967), "A multiproportions randomized response model," *J. Amer. Statist. Assoc.*, 62, 990-1008.

Anderson, H. (1976), "Estimation of a proportion through randomized response," *Int. Stat. Rev.* 44, 213-217.

Bourke, P.D. (1981), "On the analysis of some multivariate randomized response designs for categorical data," *J. Statist. Plann. Inference* 5, 165-170.

Bourke, P.D. (1982), "Randomized response multivariate designs for categorical data," *Commun. Statist. Theory Methods* 11, 2889-2901.

Bourke, P.D., and Dalenius, T. (1976), "Some new ideas in the realm of randomized inquires," *Int. Stat. Rev.* 44, 219-221.

Bourke, P.D. (1981), "On the analysis of some multivariate randomized response designs for categorical data," *J. Statist. Plann. Inference* 5, 165-170.

Bradburn, N.M., and Sudman, S. (1979), *Improving interview method and questionnaire design,* San Francisco: Jossey-Bass.

Campbell, C., and Joiner, B.L. (1973), "How to get the answer being you've asked the question," *The American Statistician*, 27(5), 229-231.

Chaudhuri, A. and Mukerjee, R.(1988), *Randomized Response: Theory and Techniques,* New York: Marcel Dekker.

Chaudhuri, A. (2001), "Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population," *J. Statist. Plann. Inference* 94, 37-42.

Cochran, W.G. (1977), *Sampling Techniques,* 3rd edn. New York: John Wiley and Sons.

Danermark, B., and Swensson, B. (1987), "Measuring drug use among Swedish adolescents: randomized response versus anonymous questionnaires," *J. Off. Statist.,* 3, 439-448.

Deming, W.E. (1960). *Sample Design in Business Research.* John Wiley & Sons, New York

Duffy, J.C., and Waterton, J.J. (1988), "Randomized response vs. direct questioning: estimating the prevalence of alcohol related problems in a field survey," *Aust. J. Statist.*, 30, 1-14.

Eichhorn, B.H., and Hayre, L.S. (1983), "Scrambled randomized methods for obtaining sensitive quantitative data," *J. Statist. Plann. Inference* 7, 307-316.

Eriksson, S.A. (1973), "A new model for randomized response," *Int. Stat. Rev.* 41, 101-113.

Flinger, M.A., Policello, G.E. II, and Singh, J. (1977), "A comparison of two randomized response survey methods with consideration for the level of respondent protection," *Commun. Statist. Theory Methods* 6, 1511-1524.

Fox, J.A., and Tracy, P.E. (1984), "Measuring associations with randomized response," *Social Science Research.* 13, 188-197.

Goodman, L.A. (1964), "Simultaneous confidence intervals for contrasts among multinomial populations," *Ann. Math. Statist.* 35, 716-725.

Greenberg, B.G., Abul-Ela, A.A., Simmons, W.R. and Horvitz, D.G. (1969), "The unrelated question randomized response: theoretical framework," *J. Amer. Statist. Assoc.*, 64, 529-539.

Greenberg, B.G., Kuebler, R.R., Abernathy, J.R. and Horvitz, D.G. (1971), "Applications of the RR technique in obtaining qualitative data," *J. Amer. Statist. Assoc.*, 66, 243-250.

Greenberg, B.G., Kuebler, R.R., Abernathy, J.R. and Horvitz, D.G. (1977), "Respondent hazards in the unrelated question randomized response model," *J. Statist. Plann. Inference* 1, 53-60.

Himmelfarb, S., and Edgell, S.E. (1980), "Additive constants model: A randomized response technique for eliminating evasiveness to quantitative response questions," *Psychological Bulletin*, 87, 525-530.

Hong, K., Yum, J., and Lee, H. (1994), "A stratified randomized response technique," *Korean J. Appl. Statis.* 7, 141-147.

Horvitz, D.G., Shah, B.V., Simmons, W.R. (1967), "The unrelated question randomized response model," *Proceedings of Social Stat. Sec. Amer. Statist. Assoc.*, 65-72.

Kerkvliet, J. (1994), "Estimating a logit model with randomized data: the case of cocaine use," *Aust. J. Statist.*, 36, 9-20.

Kim, J-I., and Flueck, J.A. (1978), "An additive randomized response model," *Proceedings of Social Stat. Sec. Amer. Statist. Assoc.,* 351-355.

Kish, Leslie. (1965), *Survey Sampling*, John Wiley and Sons, New York, New York.

Kuk, A.Y.C. (1990), "Asking sensitive question indirectly," *Biometrika* 77, 436-438

Lanke, J. (1975), "On the choice of the unrelated question in Simmons' version of randomized response," *J. Amer. Statist. Assoc.* 66, 627-629.

Lanke, J. (1976), "On the degree of protection in randomized interview," *Int. Stat. Rev.* 44, 80-83.

Leysieffer, F.W., and Warner, S.L. (1976), "Respondent jeopardy and optimal designs in randomized response model," *J. Amer. Statist. Assoc.* 71, 649-656.

Liu, P.T., and Chow, L.P. (1976), "A new discrete quantitative randomized response models for quantitative data," *J. Amer. Statist. Assoc.* 71, 72-73.

Ljungqvist, L. (1993), "A unified approach to measures of privacy in randomized response models: A utilitarian perspective," *J. Amer. Statist. Assoc.* 88, 97-103.

Mangat, N.S. (1994), "An improved randomized response strategy," *J. Roy. Statist. Soc. Ser. B* 56 (1), 93-95.

Mangat, N.S. and Singh, R. (1990), "An alternative randomized response procedure," *Biometrika* 77, 439-442.

Mangat, N.S., Singh, R., and Singh, S., Singh, B. (1993), "On Moors' randomized response model," *Biom. J.* 35 (6), 727-732.

Mangat, N.S., Singh, R., and Singh, S. (1997), "Violation of respondent's privacy in Moors' model- its rectification through a random group strategy response model," *Commun. Statist. Theory Methods* 26 (3), 243-255.

Moors, J.J.A. (1971), "Optimization of the unrelated question randomized response model," *J. Amer. Statist. Assoc.* 66, 627-629.

Mukhopadhyay, P. (1980), "On estimation of some confidential characters from survey data," *Bull. Calcutta. Statist. Assoc.* 29, 133-141.

Nayak, T.K. (1994), "On randomized response surveys for estimating a proportion," *Commun. Statist. Theory Methods* 23 (11), 3303-3321.

O'Hagan, A. (1987), "Bayes Linear Estimators For Randomized Response Models," *J. Amer. Statist. Assoc.* 82, 580-585.

Pitz, G.F. (1980), "Bayesian Analysis of Randomized Response Models," *J. Psychological Bulletin* 87, 209-212.

Pollock, K.H., and Beck, Y. (1976), "A comparison of three randomized response models for quantitative data," *J. Amer. Statist. Assoc.* 71, 884-886.

Singh, S., Singh, R., and Mangat, N.S. (2000), "Some alternative strategies to moors' model in randomized response model," *J. Statist. Plann. Inference* 83, 243-255.

Scheers, N.J. (1992), "A review of randomized response techniques," *Measurement and evaluation in counseling and development* 25, April, 27-41

Tracy, D.S. and Mangat, N.S. (1996), “On respondent’s jeopardy in two alternate questions randomized response model,” *J. Statist. Plann. Inference* 55, 107-114.

Tracy, P.E., and Fox, J.A. (1981), “The validity of randomized response for sensitive measurements,” *American Sociological Review* 46, 187-200.

Unnikrishnan, N.K. and Kunte, S. (1999), “Bayesian Analysis For Randomized response Models,” *Sankhya: The Indian Journal of Statistics* 61, Series B, Pt 3, 422-432.

Warner, S.L. (1965), “Randomized response: a survey technique for eliminating evasive answer bias,” *J. Amer. Statist. Assoc.* 60, 63-69.

Winkler, R.L. and Franklin, L.A. (1979), “Warner's Randomized response Model: A Baysian Approach,” *J. Amer. Statist. Assoc.* 74, 207-214.

APPENDIX

TABLE 3.3.

The Relative Efficiency of $MSE(\hat{\pi}'_{ms}) / MSE(\hat{\pi}'_{s})$.

For $\pi_s = 0.1$ such that $\pi_{s_1} = 0.08$, $\pi_{s_2} = 0.13$, $w_1 = 0.6$ and $w_2 = 0.4$

				P							
M	n	T	T_r	0.09	0.12	0.15	0.18	0.21	0.24	0.27	0.3
0.3	100	0.9	0.9	14.9033	16.3782	19.2583	24.924	37.8001	79.488	504.855	445.089
		0.9	0.8	13.806	15.4362	18.3895	24.0418	36.7535	77.7779	496.508	439.525
		0.9	0.7	12.2865	14.0746	17.0939	22.6969	35.1352	75.1151	483.508	430.919
		0.8	0.9	14.8983	16.383	19.2755	24.9611	37.8786	79.6998	506.498	446.802
		0.8	0.8	13.608	15.242	18.1897	23.8206	36.4752	77.3137	494.333	438.295
		0.8	0.7	11.954	13.7304	16.7198	22.2576	34.5434	74.0377	477.778	426.89
		0.7	0.9	15.0049	16.5086	19.4332	25.1784	38.2285	80.4797	511.74	451.688
		0.7	0.8	13.5111	15.1599	18.1221	23.7712	36.4584	77.4018	495.683	440.19
		0.7	0.7	11.7074	13.4844	16.4646	21.9764	34.1965	73.4848	475.438	425.897
0.3	500	0.9	0.9	12.7037	14.4189	17.3923	22.9773	35.4427	75.5921	485.892	432.787
		0.9	0.8	9.17993	10.9691	13.8324	18.993	30.297	66.5341	438.68	399.439
		0.9	0.7	7.02853	8.59921	11.1095	15.6337	25.562	57.5283	388.498	361.979
		0.8	0.9	12.5427	14.3065	17.3398	23.0159	35.6667	76.4187	493.445	441.516
		0.8	0.8	8.49684	10.2425	13.0317	18.0561	29.0677	64.4295	428.808	394.17
		0.8	0.7	6.24512	7.7	10.034	14.2558	23.554	53.614	366.503	345.938
		0.7	0.9	12.855	14.7257	17.923	23.8886	37.1709	79.9665	518.459	465.798
		0.7	0.8	8.11689	9.88196	12.6977	17.7667	28.8817	64.6387	434.342	403.064
		0.7	0.7	5.64971	7.04007	9.27965	13.3464	22.3391	51.5436	357.34	342.186
0.3	1000	0.9	0.9	10.7713	12.5753	15.5354	20.9484	32.8893	71.2344	464.105	418.335
		0.9	0.8	6.89792	8.45062	10.9328	15.408	25.2333	56.888	384.921	359.432
		0.9	0.7	5.35671	6.58655	8.57586	12.202	20.2441	46.4072	320.467	306.456
		0.8	0.9	10.4733	12.3528	15.4135	20.9884	33.2709	72.7487	478.448	435.307
		0.8	0.8	5.97555	7.42373	9.74985	13.9631	23.2582	53.376	367.894	350.134
		0.8	0.7	4.42992	5.48332	7.20348	10.3679	17.4494	40.699	286.825	280.741
		0.7	0.9	10.9662	13.0483	16.4202	22.5443	36.0254	79.3924	526.178	482.373
		0.7	0.8	5.45595	6.90628	9.24624	13.5031	22.9393	53.6892	377.317	366.012
		0.7	0.7	3.72361	4.67124	6.23777	9.15339	15.7524	37.6689	272.775	274.766
0.3	2000	0.9	0.9	8.34561	10.0803	12.8505	17.8391	28.7716	63.8867	425.921	392.151
		0.9	0.8	5.08696	6.27613	8.2041	11.726	19.5534	45.0749	313.15	301.385
		0.9	0.7	4.28556	5.2154	6.72805	9.50277	15.6979	35.99	250.095	242.593
		0.8	0.9	7.87559	9.70875	12.6284	17.8815	29.4073	66.5605	452.165	424.057
		0.8	0.8	3.97469	4.99	6.66136	9.75911	16.7417	39.8394	286.571	286.239
		0.8	0.7	3.26691	3.97316	5.13911	7.30996	12.2306	28.6014	204.405	205.75
		0.7	0.9	8.59533	10.7781	14.2473	20.4842	34.1781	78.4245	539.706	512.402
		0.7	0.8	3.34426	4.33706	5.9982	9.1241	16.2738	40.2798	301.187	312.25
		0.7	0.7	2.48955	3.05745	4.01926	5.85544	10.1214	24.6724	185.301	197.218

For $\pi_s = 0.2$ such that $\pi_{s_1} = 0.18$, $\pi_{s_2} = 0.23$, $w_1 = 0.6$ and $w_2 = 0.4$.

				P							
M	n	T	T_r	0.09	0.12	0.15	0.18	0.21	0.24	0.27	0.3
0.3	100	0.9	0.9	10.5836	12.3917	15.3462	20.7374	32.619	70.7659	461.731	416.741
		0.9	0.8	8.61811	10.3726	13.1733	18.2144	29.2537	64.6719	429.176	393.255
		0.9	0.7	7.05313	8.63519	11.1629	15.7182	25.715	57.9076	391.321	364.894
		0.8	0.9	10.499	12.3358	15.3297	20.7858	32.8057	71.4102	467.495	423.355
		0.8	0.8	8.11826	9.83661	12.578	17.5121	28.3243	63.0649	421.538	389.076
		0.8	0.7	6.40599	7.89535	10.2816	14.5932	24.0812	54.7309	373.491	351.871
		0.7	0.9	10.722	12.6379	15.7547	21.4287	33.9251	74.0742	486.426	441.853
		0.7	0.8	7.84683	9.57839	12.339	17.3073	28.2008	63.2548	425.918	395.989
		0.7	0.7	5.91662	7.35531	9.66696	13.8554	23.0992	53.0614	366.095	348.788
0.3	500	0.9	0.9	7.31231	8.95134	11.5659	16.273	26.5978	59.8395	404.074	376.66
		0.9	0.8	4.84503	5.98172	7.82918	11.2124	18.7503	43.3887	302.904	293.257
		0.9	0.7	4.18937	5.09623	6.57279	9.2839	15.3432	35.2127	245.129	238.429
		0.8	0.9	6.91835	8.63625	11.3786	16.3228	27.1964	62.3559	428.974	407.221
		0.8	0.8	3.80273	4.77222	6.37261	9.34716	16.0705	38.3704	277.259	278.517
		0.8	0.7	3.20202	3.891	5.02943	7.15109	11.9652	27.9998	200.411	202.252
		0.7	0.9	7.54533	9.58109	12.8303	18.6915	31.6007	73.4582	511.933	491.903
		0.7	0.8	3.21377	4.16047	5.7495	8.74919	15.6312	38.8053	291.407	303.759
		0.7	0.7	2.44905	3.00321	3.94256	5.73757	9.91224	24.1676	181.726	193.861
0.3	1000	0.9	0.9	5.42448	6.77228	8.94798	12.9066	21.6761	50.2064	349.551	336.257
		0.9	0.8	3.78861	4.61663	5.97293	8.47805	14.1102	32.6907	230.337	227.343
		0.9	0.7	3.62866	4.34496	5.50574	7.62742	12.3485	27.7593	189.676	182.122
		0.8	0.9	4.85197	6.29303	8.64233	12.9575	22.6114	54.3733	392.547	390.957
		0.8	0.8	2.59443	3.19776	4.21722	6.1586	10.657	25.9577	194.344	205.642
		0.8	0.7	2.5747	3.04093	3.04093	5.23515	8.4669	19.2209	134.759	135.635
		0.7	0.9	5.7121	7.64492	10.8051	16.6274	29.7008	72.915	536.055	542.355
		0.7	0.8	1.91656	2.47611	3.46069	5.40714	10.0782	26.5157	214.105	242.965
		0.7	0.7	1.77012	2.07931	2.61182	3.64766	6.10469	14.6783	111.791	124.88
0.3	2000	0.9	0.9	3.76852	4.72862	6.31445	9.26366	15.9335	38.0688	275.334	276.909
		0.9	0.8	3.14316	3.74968	4.73948	6.56209	10.6496	24.1049	166.864	163.607
		0.9	0.7	3.32099	3.92384	4.89193	6.64376	10.4986	22.9214	151.349	139.951
		0.8	0.9	3.03938	4.09542	5.8898	9.31577	17.2616	44.3152	342.962	367.066
		0.8	0.8	1.85619	2.19784	2.785	3.9244	6.6197	15.9958	121.819	135.174
		0.8	0.7	2.23047	2.56444	3.10615	4.09743	6.30589	13.5225	89.3825	85.7445
		0.7	0.9	4.10403	5.82906	8.76777	14.3939	27.484	72.2307	568.889	616.465
		0.7	0.8	1.124	1.40639	1.9398	3.06537	5.93685	16.6527	146.491	184.18
		0.7	0.7	1.39757	1.56144	1.84633	2.40663	3.75266	8.51893	63.4547	73.2192

For $\pi_s = 0.3$ such that $\pi_{s_1} = 0.28$, $\pi_{s_2} = 0.33$, $w_1 = 0.6$ and $w_2 = 0.4$.

				P							
M	*n*	*T*	T_r	0.09	0.12	0.15	0.18	0.21	0.24	0.27	0.3
0.3	100	0.9	0.9	8.24574	9.97469	12.7337	17.7001	28.5826	63.5414	424.087	390.869
		0.9	0.8	6.18657	7.63204	9.9525	14.1524	23.4081	53.3476	365.183	345.2
		0.9	0.7	5.12837	6.31768	8.24594	11.769	19.601	45.1427	313.457	301.666
		0.8	0.9	8.06168	9.83423	12.66	17.7452	28.8946	64.7677	435.828	404.964
		0.8	0.8	5.45753	6.81202	8.99768	12.973	21.777	50.4113	350.752	337.176
		0.8	0.7	4.31491	5.34844	7.03894	10.1539	17.1362	40.0989	283.656	278.807
		0.7	0.9	8.40077	10.3286	13.3988	18.9222	31.0372	70.0679	474.774	444.134
		0.7	0.8	5.05034	6.40312	8.59669	12.6051	21.5249	50.6936	358.809	350.772
		0.7	0.7	3.69613	4.63643	6.19145	9.08693	15.6433	37.4279	271.232	273.471
0.3	500	0.9	0.9	4.8508	6.07912	8.07521	11.7294	19.8703	46.4994	327.569	319.227
		0.9	0.8	3.59097	4.35867	5.61722	7.94432	13.183	30.4907	214.9	212.783
		0.9	0.7	3.53298	4.21798	5.32671	7.35064	11.8481	26.5069	180.238	172.334
		0.8	0.9	4.28865	5.60286	7.76683	11.7779	20.8259	50.8243	372.831	377.591
		0.8	0.8	2.40637	2.94588	3.86127	5.61231	9.68927	23.6275	177.885	190.236
		0.8	0.7	2.48342	2.91711	3.62988	4.95253	7.94383	17.8817	124.457	124.776
		0.7	0.9	5.12791	6.93747	9.92798	15.4917	28.094	70.0877	523.931	539.112
		0.7	0.8	1.73359	2.22689	3.10404	4.85591	9.10227	24.1935	198.196	229.032
		0.7	0.7	1.68211	1.9577	2.43336	3.36093	5.56749	13.2923	101.125	113.776
0.3	1000	0.9	0.9	3.42849	4.29262	5.72856	8.41538	14.5288	34.9392	255.108	259.785
		0.9	0.8	3.03932	3.61069	4.5418	6.25393	10.0882	22.6914	156.175	152.53
		0.9	0.7	3.27207	3.85812	4.7978	6.49528	10.223	22.2072	145.709	133.714
		0.8	0.9	2.70794	3.66236	5.30196	8.46588	15.879	41.366	325.535	354.879
		0.8	0.8	1.7579	2.06246	2.58713	3.60808	6.03075	14.4883	110.335	123.375
		0.8	0.7	2.1839	2.5004	3.01222	3.94559	6.01701	12.7557	83.1931	78.7813
		0.7	0.9	3.75674	5.38231	8.17959	13.5877	26.2895	70.1012	560.836	617.917
		0.7	0.8	1.02867	1.2726	1.74153	2.74591	5.34241	15.151	135.436	173.637
		0.7	0.7	1.35274	1.49859	1.75247	2.25244	3.4555	7.72372	57.034	66.09
0.3	2000	0.9	0.9	2.39672	2.93289	3.84281	5.58378	9.63836	23.5039	177.003	189.401
		0.9	0.8	2.73	3.1826	3.91076	5.23112	8.14287	17.5479	115.015	106.868
		0.9	0.7	3.13479	3.66682	4.51303	6.02726	9.31519	19.7394	125.159	109.561
		0.8	0.9	1.56127	2.18542	3.32121	5.63592	11.3498	32.0099	274.555	327.986
		0.8	0.8	1.39429	1.55684	1.83948	2.39539	3.73102	8.46114	62.9895	72.7049
		0.8	0.7	2.02631	2.2789	2.67967	3.39462	4.94066	9.81363	58.6356	50.0153
		0.7	0.9	2.76207	4.19865	6.77458	11.9608	24.6374	70.1145	600.616	711.228
		0.7	0.8	**	**	**	1.46921	2.97897	9.18777	91.4463	131.657
		0.7	0.7	1.17944	1.25455	1.38587	1.64591	2.27572	4.52761	30.794	36.2661

* * does not satisfy the condition (3.5.1).

For $\pi_S = 0.4$ such that $\pi_{S_1} = 0.38$, $\pi_{S_2} = 0.43$, $w_1 = 0.6$ and $w_2 = 0.4$.

				P							
M	n	T	T_r	0.09	0.12	0.15	0.18	0.21	0.24	0.27	0.3
0.3	100	0.9	0.9	6.78214	8.35519	10.8685	15.4	25.3552	57.471	391.007	367.209
		0.9	0.8	4.91572	6.08281	7.97882	11.4492	19.1777	44.4298	310.353	300.438
		0.9	0.7	4.27934	5.22619	6.76968	9.60663	15.953	36.7829	257.093	250.755
		0.8	0.9	6.49248	8.12044	10.7259	15.4332	25.8023	59.3808	410.111	390.865
		0.8	0.8	4.02209	5.0503	6.7414	9.87316	16.9269	40.2431	289.116	288.337
		0.8	0.7	3.37099	4.12251	5.36409	7.67681	12.9202	30.3667	217.752	219.343
		0.7	0.9	6.94502	8.81062	11.7983	17.2013	29.123	67.8301	473.797	456.391
		0.7	0.8	3.51638	4.52713	6.2108	9.36619	16.5554	40.6009	300.814	309.088
		0.7	0.7	2.67804	3.30924	4.37386	6.3973	11.0762	26.9563	201.309	212.063
0.3	500	0.9	0.9	3.56898	4.47356	5.97288	8.77097	15.121	36.2669	263.748	267.153
		0.9	0.8	3.092	3.68518	4.65361	6.43773	10.4411	23.6275	163.636	160.691
		0.9	0.7	3.29419	3.89034	4.84776	6.58046	10.3937	22.6841	149.765	138.543
		0.8	0.9	2.88958	3.88029	5.57123	8.81409	16.3676	42.1899	328.395	354.017
		0.8	0.8	1.836	2.1701	2.74455	3.85992	6.50004	15.6907	119.507	132.813
		0.8	0.7	2.2165	2.54685	3.08275	4.06353	6.24892	13.3914	88.4889	84.9292
		0.7	0.9	3.87228	5.48872	8.25535	13.5765	26.0119	68.6977	544.428	594.262
		0.7	0.8	1.12088	1.39668	1.91804	3.01895	5.82985	16.3287	143.657	180.897
		0.7	0.7	1.39326	1.55541	1.83735	2.3919	3.72433	8.4433	62.8459	72.5461
0.3	1000	0.9	0.9	2.51595	3.09292	4.06947	5.93258	10.2581	25.0016	187.627	199.395
		0.9	0.8	2.76084	3.22681	3.97829	5.34463	8.3671	18.1652	120.179	112.895
		0.9	0.7	3.14673	3.68432	4.54038	6.0744	9.41101	20.0124	127.544	112.509
		0.8	0.9	1.70883	2.37213	3.5671	5.97997	11.8845	33.0567	279.516	329.064
		0.8	0.8	1.43902	1.61946	1.93283	2.54832	4.02476	9.24369	69.2699	79.6205
		0.8	0.7	2.04369	2.30383	2.71789	3.45927	5.06972	10.1746	61.7258	53.743
		0.7	0.9	2.86528	4.30717	6.87787	12.0244	24.5336	69.1586	586.676	687.627
		0.7	0.8	**	**	1.04621	1.6345	3.28369	9.95526	97.1088	137.053
		0.7	0.7	1.20094	1.28492	1.43167	1.72207	2.42483	4.93503	34.1781	40.1759
0.3	2000	0.9	0.9	1.83303	2.16601	2.73857	3.85036	6.48226	15.6453	119.162	132.46
		0.9	0.8	2.58244	2.97678	3.6042	4.72729	7.16652	14.9004	93.0833	81.3431
		0.9	0.7	3.07054	3.57724	4.37942	5.80684	8.88498	18.5582	115.172	97.5411
		0.8	0.9	0.94308	1.35961	2.16576	3.90087	8.4036	25.4712	235.552	304.413
		0.8	0.8	1.22518	1.31912	1.48318	1.80758	2.5919	5.3903	37.946	44.5067
		0.8	0.7	1.95441	2.17753	2.52683	3.13981	4.43848	8.42374	46.825	35.8139
		0.7	0.9	2.2122	3.51392	5.9147	10.8859	23.3857	69.5414	624.675	779.858
		0.7	0.8	**	**	**	**	1.80979	6.14586	68.0847	108.11
		0.7	0.7	1.10159	1.14433	1.21924	1.36794	1.72919	3.02558	18.2168	21.5661

* * does not satisfy the condition (3.5.1).

2

VITA

Jong-Min Kim

Candidate for the Degree of

Doctor of Philosophy

Thesis: NEW APPROACHES TO RANDOMIZED RESPONSE TECHNIQUE

Major Field: Statistics

Biographical:

Personal Data: Born in Pyongtaek, South Korea, on February 05, 1972.

Education: Graduated from Pyongtaek High School, Pyongtaek, South Korea, in February, 1990; Received a Bachelor of Science Degree with a Major in Mathematics Education from Chongju University, Chongju, South Korea, in February, 1994; Received the Master of Science Degree with a Major in Mathematics from Chung-Ang University, Seoul, South Korea, in February, 1996. Completed the requirements for the Doctor of Philosophy Degree with a Major in Statistics at Oklahoma State University in May 2002.

Experience: Graduate Assistant, Graduate College, Chung-Ang University, August, 1994, to July, 1995. Teaching Assistant, Department of Mathematics, Oklahoma State University, August, 1996, to August, 1998. Teaching Assistant, Department of Statistics, Oklahoma State University, January, 1999, to present.

Professional Memberships: American Statistics Association; Korean-American Scientists and Engineers Association