

DIS
STAT
U&I

**USING RANDOMIZED RESPONSE TO ESTIMATE
THE TRUTHFUL REPLIES**

By

AYESHA NAZUK

DEPARTMENT OF STATISTICS

QUAID-I-AZAM UNIVERSITY

ISLAMABAD, PAKISTAN

2007

USING RANDOMIZED RESPONSE TO ESTIMATE
THE TRUTHFUL REPLIES

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

*In the name of Allah most Gracious Most
Merciful*

USING RANDOMIZED RESPONSE TO ESTIMATE THE TRUTHFUL REPLIES

By

Ayesha Nazuk

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF THE MASTER OF PHILOSOPHY**

IN

STATISTICS

AT THE FACULTY OF SCIENCES

DEPARTMENT OF STATISTICS

QUAID E AZAM UNIVERSITY

ISLAMABAD, PAKISTAN.

DEPARTMENT OF STATISTICS

QUAID E AZAM UNIVERSITY

ISLAMABAD, PAKISTAN.

(2007)

USING RANDOMIZED RESPONSE TO ESTIMATE THE TRUTHFUL REPLIES

FORWARDING SHEET

The thesis entitled “USING RANDOMIZED RESPONSE TO ESTIMATE THE TRUTHFUL REPLIES” Submitted by Ms. Ayesha Nazuk in partial fulfillment of the requirements for the M.Phil degree in Statistics has been completed under my guidance and supervision. I am entirely satisfied with the quality of the research work done by the student.

Signature 


(Dr. Javid Shabbir)
Associate Professor

USING RANDOMIZED RESPONSE TO ESTIMATE THE TRUTHFUL REPLIES

DECLARATION

I Ayesha Nazuk d/o Rao Shamshad Ahmed Khan Registration No. 5535-ST/MP-2004, a student of Master of Philosophy in Statistics at Quaid-i-Azam University Islamabad, Pakistan, do hereby solemnly declare that the thesis entitled “USING RANDOMIZED RESPONSE TO ESTIMATE THE TRUTHFUL REPLIES” submitted by me in partial fulfillment of the requirements for Master of Philosophy in Statistics is my original work and has not been submitted and shall not, in future, be submitted by me for obtaining any degree from this or any other University or Institution.

Date: 17/02/07

Signature 
(Ayesha Nazuk)

Certificate


Using Randomized Response to Estimate the Truthful Replies.


by

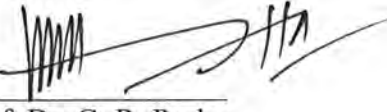
Ayesha Nazuk

A thesis submitted in partial fulfillment of the requirements for the degree of
the Master of Philosophy in Statistics.

We accept this thesis as conforming to the required standard.

1. 
Prof. Dr. Muhammad Aslam,
(Chairman)

2. 
Dr. Javid Shabbir (Assoc. Prof),
(Supervisor)

3. 
Prof. Dr. G. R. Pasha
(External Examiner)

Dated: 17/02/07

Department of Statistics,
Quaid-i-Azam University, Islamabad,
Pakistan.
Year: 2007

Dedicated to

My teachers, my parents, my brother Rao Shahbaz Ali Khan, all my friends, especially Sadia Nadir and my students.

ACKNOWLEDGEMENTS

Knowledge is a pursuit for the truth which demands a lot of adeptness and efforts to yield a fruitful addition to the never-ending stream of information. Nobody but **Almighty Allah** can help all the humanity in discovering and inventing new strategies and techniques.

During my research work I faced a lot of twisting moments where I felt very much chaotic, but it was all due to the kindness of **Almighty Allah** to hold me upright and resolute. I am really indebted to my kind supervisor Assistant Professor **Dr. Javed Shabbir** who kept on encouraging me and brought me out of every pitfall I faced. It has been a pleasure working with him and I felt very much convenient discussing and even arguing with him sometimes.

I would like to present my gratitude to **Dr. Muhammad Aslam**, Chairman of the Department of Statistics, Q.A.U. It has been his admirable kindness to facilitate me in doing my research work with all conveniences. I would like to thank all my teachers who proved like mentors to me. I am especially thankful to my teacher, **Mr. Zawar Hussain**, in giving me fruitful ideas for pursuing further in my research.

I would like to express my thanks to my **parents** and my **siblings** especially **Mr. Rao Shahbaz Ali Khan**, whose support always lifted my spirit up.

I am really thankful to Professor **Dr. Faqir Muhammad** at AIOU, in allowing me to spend a proportionate share of my time in accomplishing my thesis even when I had to fulfill the departmental duties.

I am also obliged to **Ms. Sadia Nadir** who is really a bounty for me from heavens and who is always there encouraging and appreciating my efforts.

In the end I would like to thank all my friends, all my seniors, students and colleagues in making me a well-balanced personality and backing me up all the times during my research work.

Signature _____
(Ayesha Nazuk)

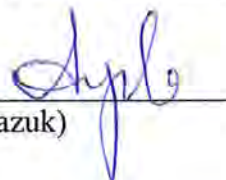


TABLE OF CONTENTS

- ❖ PROLOUGE
- ❖ ABSTRACT

1	INTRODUCTION	1-5
	1.1 Survey Sampling	1
	1.2 Sensitive Information	2
	1.3 Dealing with Sensitive Information	2
	1.4 Various Situations	3
2	REVIEW OF LITERATURE	6-33
3	A NEW MIXED RANDOMIZED RESPONSE MODEL	34-42
	3.1 Objectives of the Proposed Model	34
	3.2.1 The Proposed Model	34
	3.2.2 Testing the Hypothesis of Truthful Reporting	38
	3.2.3 Estimating the True Probability of Truthful Reporting	41
4	A NEW MIXED STRATIFIED RANDOMIZED RESPONSE MODEL	43-48
	4.1 The Need for Stratification	43
	4.2 Stratified Proposed Model	43
5	POPULATION BASED VIRTUAL RESULTS	49-56
	5.1 Introduction	49

5.2 Descriptive Statistics of the Data Sets Considered, under Truthful Reporting Atmosphere	49
5.3 Efficiency Comparison of the Proposed Model with Competing Models	49
5.4 Comparison of the Data sets Considered, under Stratified Random Sampling, Truthful Reporting Atmosphere	53
5.5.1 Real Life Survey in Simple Random Sampling Scheme	54
6 CONCLUSIONS AND RECOMMENDATIONS	57
LITERATURE CITED	58-60
APPENDIX	61-69

PROLOUGE

The foundation of Statistics is erected on estimation and inferences. The convoluted difference between Mathematics and Statistics is that the later is unavoidably inductive in nature, while the former may be inductive, deductive or contradictive.

Assortment of courses of action have been devised in estimation of the unidentified parameters of the innocuous variables but the estimation and inferences of the sensitive variables is somewhat new field. Oodles of gratefulness is owed to Warner (1965) who lead the way to the technique of randomized response.

This thesis is a diminutive effort of improving the estimation of population proportion of the sensitive qualitative variables.

Chapter one introduces the readers of the sensitive study variables and expresses their indispensability. Chapter two evolves around a comprehensive review of the literature, with efforts not to ignore any one who worked on sensitive qualitative variables. Of course it is not possible to enlist and discuss all the researchers and statisticians who worked on the mentioned field.

Chapter three and four presents a new proposed model to increase the efficiency of estimation of population proportion of the sensitive qualitative variables. In third chapter simple random sampling and in fourth chapter stratified random sampling scheme has been incorporated.

Chapter five supports the proposed model by presenting numerical and real life survey based results.

Chapter six aims to help all those who want to pursue further in this direction and who want to have a finale-look at the work done in previous chapters.

ABSTRACT

Variety of work has been done in estimation of the unknown parameters of the innocuous variables but the estimation and inferences of the sensitive variables is relatively new field. Lots of gratitude is due to Warner (1965) who pioneered the technique of randomized response.

This thesis contributes to eliminate the evasive answer bias while dealing with the qualitative sensitive study variables. We have worked on the Kim and Warde (2004) model to improve its methodology and proposing a new mixed randomized response model for estimation of population proportion of the qualitative sensitive variables. We have introduced a variant in the Kim and Warde (2004) model and increased the efficiency of estimation of population proportion of the qualitative sensitive variables. The proposed model is also extended in the stratified random sampling scheme, and is proved to be efficient in this atmosphere as well.

Chapter 1

INTRODUCTION

1.1 Survey Sampling

Most of the phenomenon are subject to random variations, and nothing could be said with surety about these random phenomenon. To deal such random phenomenon, we have the field of statistics that erects its foundation stone on the aim of studying random phenomenon.

The very first step in a survey sampling is collection of reliable data. Question arises why we don't use archival records to validate, countercheck, erect or swerve from some theory or hypothesis concerning human populations. The answer carries a host of argument but most vibrant one are:

- i. That archival data are sometimes missing for a particular variable of interest such as: in criminology undetected or unreported crimes are not filed, similarly, in health-sciences data about patients, who did not confer/treated by any doctor or health expert remains missing from hospital's record.
- ii. That archival data are, sometimes, inaccessible for a particular variable of interest such as: it is legal right of criminals that there personal and crime record cannot be handed over to anyone/researchers unless he/she has some legal permission to access these data. Sometimes, it is a cumbersome task to get a permission to access these data for usage in research.

Therefore, survey methods are an apposite substitute to study any human populations. However, one must keep it in mind that survey methods are not free from shortcomings and entail a handsome degree of proficiency in data collection,

compilation, elucidation, interpretation and construal etc. Therefore one must get him/her properly trained in survey methods before endeavoring to use them.

1.2 Sensitive Information

If a variable or a characteristic of interest is so that, its presence or absence causes the person to be socially undesirable, its presence or absence is perceived illegal, it is considered as a stigma, or its revelation concerns the privacy, then that variable or characteristic may be referred to be as sensitive. In survey methodology whenever we have to deal with a study variable that is sensitive in nature, there arises the problem of obtaining reliable data about it. A variable is also termed as sensitive or stigmatized if it is subject to social desirability response bias (SDB). SDB is defined as the bias, which is introduced when the respondents conceal the true response due to the fear that its disclosure will make them socially undesirable (see Gupta and Thornton, 2002).

There are basically two types of sensitive variables (i) Quantitative and (ii) Qualitative (see Section 1.4). It is also noteworthy that survey methodologies devised for innocuous variables prove inefficacious in dealing with sensitive or stigmatized variables. A variable is said to be innocuous if respondents feel that disclosure of information on that variable will not cause any concern to their privacy.

1.3 Dealing with Sensitive Variables

We know that statistics evolves around the motive of analyzing about the parameters of the concerned distribution. However, it is noticeable that the usual survey methods fail to estimate the parameters of the distribution of a sensitive variable. Because of the presence of substantial bias in sensitive studies there is

always a great chance of concluding fuzzy and untrue results, if the techniques of innocuous variables are applied in such studies. It is not necessarily that the data collectors and interviewer are indolent on their part. But it may so, that the information required is so sensitive that no matter how adroit the interviewer is but sensitivity can hinder the true response. Due to shortcomings and previous experiences that show inefficacies of legal, administrative, anonymity methods, it was strongly felt that there is a need for some other methods that raise the general response rate and lessen the response bias. It was felt that the new strategy should have the potential to ensure that:

- i. unlike anonymous surveys, data from multiple sources may be interlinked.
- ii. unlike telephone or mail methods, allows the personal interaction of the interviewers with the interviewees.
- iii. unlike conventional survey methods, should be able to lessen or diminish the concealing of true information on the part of the interviewees.
- iv. unlike the conventional justifications, should ensure the respondents' privacy in a strongly provable and perceivable method.
- v. unlike the administrative strategies, should cease the forced disclosures.

One of techniques fulfilling these conditions is randomized response strategy pioneered by Warner (1965).

1.4 Various Situations

There may be numerous situations where we may be interested in gaining sensitive information. But we present some of the possible practical situations where sensitive study variables are involved.

- i. Suppose we are interested to estimate the average income of the head of the household in a particular locality. We know that generally people are prone to hide their income, especially, from an interviewer, who is a stranger to them and it is not easy to prepare all the respondents to confide their true income with the interviewer. Therefore income is a quantitative sensitive variable. In such a situation it is vivid that most of the respondents will underreport their income.
- ii. Suppose we are interested to estimate the proportion of alcoholic beverages users in a particular Muslim community. Since drinking alcoholic beverages is socially undesirable and religiously prohibited in a Muslim community, so respondents will not entrust their true response. In other words if they are asked the question, "Do you take alcoholic beverages?" The response will generally be "No". Here drinking status is a qualitative sensitive variable.
- iii. Suppose we are interested to estimate the proportion of persons who do not pay income tax. With the fear of accountability by the government and being socially undesirable nonpayer, respondents will never like to tell that they are involved in income tax nonpayment. So number of persons who are involved in income tax nonpayment is a qualitative sensitive variable.
- iv. Suppose, we are interested to estimate the average amount of adulteration in milk packs of a particular company. Now if the company is involved in milk adulteration, it will never like to tell that they adulterate their milk packs. The reasons may be because they do not want to lose their customers or they do not want to get penalized by concerned authorities etc. So amount of adulteration in milk packs of a particular company is a quantitative sensitive variable.

- v. Consider that a medical officer is interested to know the proportion of victims of HIV-AIDS in a particular society, which is an admixture of persons from different religions. Further, it is concluded from past experience that a Muslim has a lesser chance of being a victim of HIV-AIDS (by virtue of being sexually restrained). Therefore the sensitive question, “Are you a victim of HIV-AIDS?” is a function of the non-sensitive question “Are you Muslim”. Therefore it is needed that the medical officer stratifies the population into two strata, one labeled “Muslim” and the other “Non-Muslim”. In this situation if simple random sampling is employed it may yield a sample with majority of sampled individuals from either of the two strata yielding an unrepresentative sample.

Hence in all of the above examples/cases, the parameters of the distribution of study variable will not be estimated reliably; unless some other survey methodology is applied that has the potential to deal with sensitive study variables.

By the virtue of many researchers and statisticians we have different techniques of dealing with sensitive variables including *the randomized response method*. In the coming chapter we shall discuss some of those techniques.

It is also noteworthy that we have worked on introducing a new model to estimate the proportion of qualitative sensitive study variables. We have introduced a new variant of the Kim and Warde (2004) model and have successfully improved the efficiency of estimation.

Chapter 2

REVIEW OF LITERATURE

2.1 Dealing with Sensitive Variables

Many researchers have studied and analyzed the problem of dealing with sensitive random variables. Warner (1965) made preliminary efforts in this regard. From then onward, many efforts have been made including the efforts of Horvitz et. al. (1967), Greenberg et. al. (1969), Lanke (1975), Moors (1971), Folsom et. al. (1973), Zdep et. al. (1979), Adhikari et. al. (1984), Tracy and Fox (1987), Ljungqvist (1993), Papineau (1994), Kim and Warde (2004).

2.2 Acronyms and Notations

The following acronyms and notations have been used throughout this thesis.

N :	The population size.
n :	The sample size.
X :	The dichotomous sensitive qualitative variable.
Y :	The innocuous variable unrelated to X .
A :	The subgroup of the population possessing X .
B :	The subgroup of the population not possessing X .
π :	The population proportion of respondents possessing X .
π_y :	The population proportion of the respondents possessing Y .
$B(.)$:	The bias of the given estimator.
$V(.)$:	The variance of the given estimator.
$MSE(.)$:	The mean square error of the given estimator.
SRSWR:	simple random sampling with replacement.

2.3 Randomized Response in Completely Truthful Responding

In conventional survey methods, we often face the problems of lower response rates or even no response because of the sensitivity of the study variable. Due to these problems, many researchers worked on devising new strategies to tackle the studies involving sensitive study variables. When the population under study is believed to be

homogenous with respect to the sensitive study variable then researchers use randomized response device and obtain samples with simple random sampling procedure. Randomized response under simple random sampling has been extensively discussed by different statisticians and researchers, including Warner (1965), Kim et. al. (1978), Chang and Huang (2001) etc. Let us discuss few randomized response models.

2.3.1 Warner's Randomized Response Model

Warner (1965) is the pioneer, who considered the randomized response technique; therefore it is deemed necessary to discuss from where he initialized the concept of randomized response strategy.

Initially, Warner (1965) introduced a model for estimating the population proportion, π of dichotomous sensitive qualitative variable.

Let the dichotomous sensitive qualitative variable be represented by X . The population is dichotomized in to two mutually exclusive groups A and B , where A is the group of respondents possessing X and B is the group of respondents does not possessing X , or in other words B is the complement of A . A simple random sample of size n is drawn with replacement from the population. Each respondent is asked to use a (a randomizing device) like deck of cards or a spinner, or any other feasible randomizing device. And the spinner may point to either of the two mutually exclusive groups A and B , with probabilities p and $(1-p)$ respectively. Each respondent is requested to answer a "yes" or "no" in accordance with the situation that he/she belongs to the group pointed by the spinner. Interestingly respondents are not supposed to report that to which group the spinner pointed. In this way a privacy-preserving atmosphere is provided to the respondents, which may encourage true response rate on X . It is noteworthy that the respondents' privacy is preserved since what is reported to the interviewer is just the "yes" or "no" and not the fact that to which group respondent belongs. Assuming that all the respondents will report their true responses in above environment, the maximum likelihood estimates of the true proportion of respondents possessing the sensitive attribute are very clear-cut.

Let

p = the probability of spinner pointing to A .

n_1 = the number of respondents reporting "yes",

n_2 = the number of respondents reporting "no" = $n - n_1$

such that

$$p(X = 1) = p\pi + (1 - \pi)(1 - p)$$

and

$$p(X = 0) = (1 - \pi)p + \pi(1 - p).$$

TABLE 2.1 Sampling Scheme of Warner (1965)

Response from the Warner's Model		Statement pointed by the randomizing device used	
		$\in A$	$\notin A$
Status of the person selected in the sample	$\in A$	YES $p\pi$	NO $(1 - p)\pi$
	$\notin A$	NO $(1 - \pi)p$	YES $(1 - \pi)(1 - p)$

Then the likelihood of the sample is given by

$$L = [\pi p + (1 - \pi)(1 - p)]^{n_1} [(1 - \pi)p + \pi(1 - p)]^{n - n_1} \quad (2.3.1)$$

or

$$\log L = n_1 \log [\pi p + (1 - \pi)(1 - p)] + (n - n_1) \log [(1 - \pi)p + \pi(1 - p)]. \quad (2.3.2)$$

Now setting $\frac{\partial(\log L)}{\partial \pi} = 0$, we get the maximum likelihood estimate of

π as given below

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

The above likelihood estimate of π is estimable under the constraint that $p \neq 1/2$.

Now let us find the expected value of the estimator $\hat{\pi}$

$$\begin{aligned} E(\hat{\pi}) &= \frac{1}{2p - 1} \left[p - 1 + \left(\frac{1}{n} \right) \sum_{i=1}^n E(X_i) \right] \\ &= \frac{1}{2p - 1} [(p - 1 + \pi p) + (1 - \pi)(1 - p)] \\ &= \pi. \end{aligned} \quad (2.3.4)$$

Equation (2.3.4) shows that $\hat{\pi}$ is an unbiased estimator of the parameter π .

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

$$\begin{aligned}
 V(\hat{\pi}) &= \frac{nV(X_i)}{(2p-1)^2 n^2} \\
 &= \frac{[\pi p + (1-\pi)(1-p)][(1-\pi)p + \pi(1-p)]}{(2p-1)^2 n} \\
 &= \frac{1}{n} \left[\frac{1}{16 \left(p - \frac{1}{2}\right)^2} - \left(\pi - \frac{1}{2}\right)^2 \right]. \tag{2.3.5}
 \end{aligned}$$

We know that maximum likelihood estimators are normally distributed for appropriate sample size.

Rewriting (2.3.5) in another form, we get an unbiased estimate of $V(\hat{\pi})$ as follows

$$\hat{V}(\hat{\pi}) = \left[\frac{\left(\frac{1}{4}\right) - \left\{\hat{\pi} - \left(\frac{1}{2}\right)\right\}^2}{n} - \frac{\frac{1}{16 \left\{p - \left(\frac{1}{2}\right)\right\}^2} - \frac{1}{4}}{n} \right]. \tag{2.3.6}$$

From (2.3.6) it is clear that the variance of the estimator $\hat{\pi}$ can be easily separated in two components, one showing the variation due to sampling and other due to the use of randomizing device.

An important concern in Warner's model is the appropriate choice of n and p . Since p is the probability of pointing to the sensitive group A , therefore larger the value of p lesser protection will be offered to the respondents. But we cannot keep p very small, as doing so will make the entire procedure a close match of the conventional direct questioning method. One possible suggestion by Warner is that p should be chosen between 0.5 and 1. As far as the matter of sample size is concerned there exists an indirect relationship between n and p . Or in other words lesser the value of p greater is n required to obtain a particular degree of precision.

Warner's model is based on the assumption that respondents will report truthful response in this atmosphere. However this may not always be the case and respondents may report false responses even in this atmosphere. This is a crucial

criticism over the Warner's model. However one must not forget that it is due to his model that further researchers have embarked on the issue of surveying sensitive variables.

2.3.2 The Unrelated Question Model in Completely Truthful Reporting Atmosphere.

Greenberg et. al. (1969) discussed that in Warner's model the variance of the estimator $\hat{\pi}$ may become large if p is close to 0.5 and n is not large.

Warner considered the ratios of mean square errors of the estimators utilizing both bias and variance, to compare his suggested model with the conventional direct questioning model. This is not a realistic approach as the main aim of Warner's model was to improve the response rate itself, this remains unanswered that when is Warner's model better and applicable. For answering this question field trials or pilot surveys are must to undertake.

Warner used the values of $\pi = 0.5$ and $\pi = 0.6$ which are acceptable for less sensitive variables, like voting behaviors, but are not apt for more sensitive variables like tax evasion, illicit sexual relationships etc. if we assume that all the respondents report truthfully and, if $\pi = 0.05$ and $n = 1000$ then Warner's technique is half as efficient as the direct question method. It is only one-tenth efficient when $p = 0.20$. When only 90% of the respondents in group A are likely to admit their membership in it, the efficiency of the Warner's procedure climbs to two-thirds for $p = 0.05$, and to about one-seventh for $p = 0.2$. If as many as 50% of the members deny their membership in group A then Warner's technique becomes six times more efficient. Therefore Warner's technique requires substantial amount of lying for becoming more efficient. Since, as evasive answer bias increases Warner's technique starts gaining superiority over the direct questioning method.

In Warner's model it is assumed that the randomizing device used may point to either of the two mutually exclusive and collectively exhaustive groups. But Greenberg et. al. (1969) pointed out that according to the suggestion of Horvitz et. al. (1967), it is not necessary that the two questions/statements on the randomizing device are mutually exclusive and collectively exhaustive. Therefore, the same motive of randomizing the responses through a randomizing device can be achieved even if the two questions/statements on the randomizing device are unrelated. The unrelated question model has an additional advantage that respondent's privacy is more

protected then the Warner (1965) model. It is because of the reason that in unrelated question model the respondents do not have a fear that their answers about one statement may reveal some information about the other, because now the two statements on the randomizing device are non-mutually exclusive.

The unrelated question approach proceeds as follows:

Let the sensitive qualitative variable be represented by X . Let Y be another unrelated innocuous variable. A simple random sample of size n is drawn with replacement from the population. Each respondent is asked to use a randomizing device. And the randomizing device may choose two statements

Q (1) "I possess the sensitive trait X " and

Q (2) "I possess the innocuous trait Y "

with probabilities p and $(1-p)$ respectively.

Each respondent is requested to answer a "yes" or "no" to the question, which the randomizing device chooses. Just as in Warner's model respondents are not said to report that to which question/statement the randomizing device pointed, which is in fact a basic requirement of any randomized response strategy.

Further work is possible after deciding whether π_y is known or not.

(i) When π_y is known.

Under known π_y , the probability of a yes response in sample is given by

$$\lambda = p\pi + (1-p)\pi_y \quad (2.3.7)$$

Solving for π , we have

$$\hat{\pi}_{SIM1} = \frac{\hat{\lambda} - (1-p)\pi_y}{p}. \quad (2.3.8)$$

The variance of $\hat{\pi}_{SIM1}$ is given by

$$V(\hat{\pi}_{SIM1}) = \frac{\lambda(1-\lambda)}{np^2}. \quad (2.3.9)$$

An unbiased estimate of $V(\hat{\pi}_{SIM1})$ is

$$\hat{V}(\hat{\pi}_{SIM1}) = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)p^2}. \quad (2.3.10)$$

An interesting question arise that “can we assume that π_y is known?” Horvitz et. al. (1976) pointed out that this issue can be solved by appending an extra question in the randomizing device itself. For example

- i. The sensitive question
- ii. An instruction to say “yes”
- iii. An instruction to say “no”

with probabilities p_1, p_2 and p_3 respectively, where $\sum_{i=1}^3 p_i = 1$.

Then the probability of a “yes” response is

$$\lambda = p_1\pi + (1 - p_2)\pi_y, \quad (2.3.11)$$

where $\pi_y = \frac{p_2}{p_2 + p_3}$.

Stating (2.3.11) in another way, we get

$$\begin{aligned} \lambda &= p_1\pi + (1 - p_2) \left[\frac{p_2}{p_2 + p_3} \right] \\ &= p_1\pi + (p_2 + p_3) \left[\frac{p_2}{p_2 + p_3} \right]. \end{aligned}$$

So (2.3.11) is directly compatible with (2.3.7) with $p = p_1$ and $\pi_y = p_2 / (p_2 + p_3)$.

In this way an expression for π_y is achieved. We find similar efforts by Boruch (1972), who calls it a “contamination design”.

(ii) When π_y is unknown.

In (2.3.11) we have two unknown parameters π_y and π , Therefore one sample proves insufficient to provide the estimate of the parameters. To solve this problem we collect two random samples, using simple random sampling with replacement, of size n_1 and n_2 , such that $n = n_1 + n_2$. Then we replicate the π_y known atmosphere to the two samples.

The probability of a “yes” response in i^{th} sample ($i = 1, 2$) is given by

$$\lambda_i = p_i\pi + (1 - p_i)\pi_y. \quad (2.3.12)$$

Let n_{i1} be the number of “yes” responses in the i_{th} sample and $\hat{\lambda}_i = \frac{n_{i1}}{n_i}$. Then an unbiased estimate of π is given by

$$\hat{\pi}_{SIM2} = \frac{\hat{\lambda}_1(1-p_2) + \hat{\lambda}_2(1-p_1)}{(p_1 - p_2)}. \quad (2.3.13)$$

The observed proportions $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are binomially distributed with parameters (n_1, λ_1) and (n_2, λ_2) respectively. Therefore $\hat{\pi}_{SIM2}$ is an unbiased estimator of the parameter π , and its variance is given by

$$V(\hat{\pi}_{SIM2}) = \frac{1}{(p_1 - p_2)} \left[\frac{\lambda_1(1-\lambda_1)(1-p_2)^2}{n_1} + \frac{\lambda_2(1-\lambda_2)(1-p_1)^2}{n_2} \right]. \quad (2.3.14)$$

The unbiased estimate of $V(\hat{\pi}_{SIM2})$ is given by

$$V(\hat{\pi}_{SIM2}) = \frac{1}{(p_1 - p_2)} \left[\frac{\hat{\lambda}_1(1-\hat{\lambda}_1)(1-p_2)^2}{n_1} + \frac{\hat{\lambda}_2(1-\hat{\lambda}_2)(1-p_1)^2}{n_2} \right]. \quad (2.3.15)$$

2.3.3 Optimal Choice of the Design Parameters in Unrelated Question

Model

As we know that main concern in sensitive studies is preserving respondent's privacy subject to the matter of efficiency of estimates of population parameter(s).

Here we reproduce the recommendations of Greenberg et al. (1969) and Moors (1971) for choosing the optimal values of the design parameters n_1, n_2, p_1 and p_2 .

We skip the optimization of n because, generally $n(=n_1 + n_2)$ is fixed in advance in accordance with the available time, cost and labor etc.

First we present the issue of optimizing n_1 and n_2 :

Minimizing $V(\hat{\pi}_{SIM2})$ in equation (2.3.14) w.r.t n_1 and n_2 , subject to the constraint $n = n_1 + n_2$.

Using Cauchy-Schwartz inequality we get the optimizing condition as

$$\frac{n_1}{n_2} = \left[\frac{(1-p_2)^2 \lambda_1(1-\lambda_1)}{(1-p_1)^2 \lambda_2(1-\lambda_2)} \right]^{1/2}. \quad (2.3.16)$$

Using (2.3.16), the optimum value of $V(\hat{\pi}_{SIM2})$ is given by

$$V(\hat{\pi}_{SIM2})_{opt} = \frac{1}{(p_1 - p_2)n^{1/2}} \left[(1-p_2)[\lambda_1(1-\lambda_1)]^{1/2} + (1-p_1)[\lambda_2(1-\lambda_2)]^{1/2} \right]. \quad (2.3.17)$$

But an important task needs concern that is choosing values of λ_1 and λ_2 , since they are unknown. Possible solution is to estimate them through a pilot survey/field trial.

Now we present the optimum selection of the unrelated question Y , or π_y :

Consider that we have a function $f(\lambda) = [\lambda(1-\lambda)]^{1/2}$. Maximizing this function with respect to λ , we get $\lambda = 1/2$ as the optimum choice of λ that will maximize $f(\lambda)$. From this discussion we conclude that $[\lambda(1-\lambda)]^{1/2}$ is maximum at $\lambda = 1/2$, is symmetric about $1/2$, and is concave. In (2.3.14), π_y is involved only in λ_1 and λ_2 . So one must choose λ_1 as remote from $1/2$ as possible, or equivalently by choosing Y such that π_y is on same side of $1/2$ as is π and $|\pi_y - 1/2|$ is maximized. Anyhow if it is not vivid on which side of $1/2$, π_y lays then one possibility is to choose $\pi_y \in [0.25, 0.75]$.

Now we present the optimum selection of p_1 and p_2 :

Greenberg et. al. (1969) suggested choosing $p_1 \in [0.10, 0.20]$ or $p_1 \in [0.70, 0.90]$ and $p_2 = 1 - p_1$. However Moors (1971) showed that the precision of the estimator of π can be improved if p_1 and p_2 are chosen as apart as possible subject to the condition of privacy protection as jeopardy will be invited if p_1 and p_2 are chosen close to 1. He further mentions that p_2 should be chosen as small as possible and in fact equal to 0. This choice of p_2 will make the respondents feel comfortable since then they will be asked quite an innocuous question if $p_2 = 0$. The justification of Moors (1971) presented was that $p_2 = 0$ would make the second sample inquiry just to estimate π_y . However Mangat et al. (1997) pointed out a difficulty in the practicality of Moors (1971) model, he stated that $p_2 = 0$ means that privacy of any respondent, appearing in both independent samples and reporting “yes” in the first sample and “no” in the second, will be jeopardized. And the possibility of a respondent being selected in both samples is easily perceived and found in practice. For example in stratified sampling the inclusion of same respondent in two strata is fairly possible with large stratum size. Or even in simple random sampling it is possible for with replacement case.

Inviting further improvements in unrelated question model Mahmood et. al. (1998) pointed out that, Mangat et. al. (1997) methods need dichotomizing the population in to two random groups which is not an easy task. Moreover the variance expression of Mangat et al. (1997) model is very complicated, non-practicable and time consuming to implement. Therefore Mahmood et al. (1998) presented some new techniques which we states are free from the difficulties in Moors (1971) and Mangat et al. (1997) and fulfill the optimality conditions suggested by Moors (1971).

2.3.4 An Alternative Randomized Response Procedure

Mangat and Singh (1990) used two randomizing devices, say R_1 and R_2 , such that, each selected respondent is instructed to use randomizing device say R_1 in which there are two statements (i) “I belong to the sensitive group” and (ii) “ go to R_2 ” with probability T and $(1-T)$ respectively. On R_2 there are two statements (i) “I belong to the sensitive group” and (ii) “ I do not belong to the sensitive group” with probability p and $(1-p)$ respectively. R_2 is to be used only if directed by R_1 .

TABLE 2.2 Sampling Scheme of Mangat (1990) in Truthful response

	Randomizing device R_1		Randomizing device R_2	
Statement pointed by the chosen device with probability	$\in A$ T	Go to R_2 . $(1-T)$	$\in A$ p	$\notin A$ $(1-p)$
Respondent response	Yes $T \pi$		Yes $(1-T) \pi p$	Yes $\pi (1-p)$
				Yes $(1-\pi)(1-p)$

Therefore the probability of a “yes” response becomes

$$\theta_1 = T\pi + (1-T)[p\pi + (1-p)(1-\pi)]. \quad (2.3.18)$$

The maximum likelihood estimator of π is

$$\hat{\pi}_{MST} = \frac{\left(\frac{n^*}{n}\right) - (1-T)(1-p)}{2p-1+2T(1-p)}. \quad (2.3.19)$$

As $\left(\frac{n^*}{n}\right) \sim B(n, \theta_1^*)$, where θ_1^* is the sample counterpart of θ_1 defined in (2.3.18).

Also n^* is the number of “yes” response out of total n respondents.

The variance of $\hat{\pi}_{MST}$ is given by

$$V(\hat{\pi}_{MST}) = \frac{\pi(1-\pi)}{n} + \frac{(1-T)(1-p)\{1-(1-T)(1-p)\}}{n\{2p-1+2T(1-p)\}^2}. \quad (2.3.20)$$

An unbiased estimator of the variance of $\hat{\pi}_{MST}$ is given by

$$\hat{V}(\hat{\pi}_{MST}) = \frac{\frac{n^*}{n}\left(1 - \frac{n^*}{n}\right)}{\left[(n-1)\{2p-1+2T(1-p)\}^2\right]}. \quad (2.3.21)$$

2.4 Randomized Response in Less Than Truthful Responding

In all the above techniques/models it has been assumed that respondents will never tell a lie in the randomized response atmosphere. Whereas it may not always be true and some respondents, due to one or more of many reasons, like to conceal true response even in the randomized response atmosphere. The reason may be so that the variable is highly sensitive and some of the respondents are not fully satisfied by the protection offered by the survey methods whatsoever used. Therefore, many researchers also utilized this possibility and assumed less than completely truthful reporting in their models.

2.4.1 The Unrelated Question Model in Less Than Completely Truthful Reporting Atmosphere.

Greenberg et. al. (1969) also discussed the unrelated question model under less than completely truthful reporting. He assumed that (i) the randomizing device

has been used properly and (ii) respondents do not lie about the unrelated innocuous characteristic.

Assume that π_y is unknown and, let $T (0 \leq T \leq 1)$ be the probability that a respondents lies about their membership in group A (possessing stigmatized characteristic). Then the probability of a “yes” response in first and second sample becomes

$$\lambda_1^* = p_1 (\pi T - \pi_y) + \pi \text{ and} \tag{2.4.1}$$

$$\lambda_2^* = p_2 (\pi T - \pi_y) + \pi. \tag{2.4.2}$$

Using (2.4.1) and (2.4.2), π can easily be estimated, also its bias and variance become easily estimable. For the case when π_y is known the situation becomes simpler since then, only one sample is required.

2.4.2 An Alternative Randomized Response Procedure

Mangat and Singh (1990) developed (2.3.18) in completely truthful reporting but if respondents have a probable tendency to lie even in randomized response atmosphere then we have,

TABLE 2.3 Sampling Scheme of Mangat (1990) in Less than Completely Truthful response

	Randomizing device R_1		Randomizing device R_2	
Probability of Truth	T_1		T_2	
Statement pointed by the chosen device with probability	$\in A$ T	Go to R_2 , $(1-T)$	$\in A$ p	$\notin A$ $(1-p)$
Respondent response	Yes $T \pi T_1$		Yes $(1-T) \pi p T_2$	Yes $\pi (1-p) (1-T_2)$
				Yes $(1-\pi) (1-p)$

then probability of a “yes” response becomes

$$\theta_1 = T\pi T_1 + (1-T)[\pi P T_2 + \pi(1-p)(1-T_2) + (1-\pi)(1-p)]. \quad (2.4.3)$$

Then a biased estimate of π becomes

$$\begin{aligned} MSE(\hat{\pi}_{MSL}) &= \frac{\pi T_2(1-\pi T_2)}{n} + \frac{(1-T)(1-p)[1-(1-T)(1-p)]}{n[2p-1+2T(1-p)]^2} + \pi^2(T_2-1)^2 \\ &+ \pi T(T_1-T_2) \left[1 + \pi(n-1) \{ T(T_1-T)_2 + 4TT_2(1-p) + 2T_2(2p-1) \} \right. \\ &\left. - 2(1-T)(1-p) - 2\pi n \{ 2T(1-p) + 2p-1 \} \right] \left[n \{ 2p-1+2T(1-p) \}^2 \right]^{-1}. \end{aligned} \quad (2.4.4)$$

2.4.3 Huang et. al. Model

Huang et. al. (2005) proposed another model for detecting untruthful answering in randomized response surveys that allowed the respondents to choose the statement to answer rather than probability basis.

Two independent samples of size n_i , $i=1,2$, are drawn from the population using simple random sampling with replacement. Each respondent in first sample uses the Warner's device then, is provided two options to answer (i) report the correct response for the statement which he/she has selected, or (ii) respond to a non-sensitive question "I am a member of group Y ". Similarly, in the second sample, respondents use the Warner's device then, are provided two options to answer (i) report the correct response for the statement which he/she has selected, or (ii) respond to a non-sensitive question "I am not a member of group Y ". The proportion π_y of the innocuous trait Y is assumed to be known. Respondents report "yes" or "no", according to their true status or personal willingness. Let T be the probability of a respondent giving truthful response in both samples.

Then probability of a "yes" response in first sample is given by

$$\begin{aligned} \theta_1 &= [p\pi T + (1-p(1-\pi))] + \pi(1-T)\pi_y \text{ or} \\ \theta_1 &= (p-\pi_y)\pi T + (p+\pi_y-1)\pi + (1-p). \end{aligned} \quad (2.4.5)$$

Then probability of a "yes" response in second sample is given by

$$\theta_2 = (p+\pi_y-1)\pi T + (p-\pi_y)\pi + (1-p). \quad (2.4.6)$$

Testing for prevalence of completely truthful reporting requires setting the hypothesis structure as follows

$$H_0 : T = 1 \text{ versus } H_1 : T < 1 \quad (2.4.7)$$

or equivalently testing

$$H_0 : \theta_1 = \theta_2 \text{ versus } H_1 : \theta_1 \neq \theta_2. \quad (2.4.8)$$

So one can easily test the hypothesis that $T=1$ or not, by utilizing the fact that under the null hypothesis $H_0 : T=1$, where $\theta_i = \theta = (2p-1)\pi + (1-p)$ and $\theta_i : B(n_i, \theta)$ for $i=1, 2$. Now for large samples the critical region for testing (2.4.7) or (2.4.8) is given by

$$\frac{|\hat{\theta}_1 - \hat{\theta}_2|}{\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > z_{\alpha/2}, \text{ where } z_{\alpha/2} \text{ is the } \alpha_{th} \text{ quintile point of the standard normal}$$

distribution.

Huang et. al. (2005) constructed two different unbiased estimators for π , one under H_0 and other under H_1 . Huang et. al. used the Warner's (1965) unbiased estimator for estimating the population proportion of the sensitive variable.

The unbiased estimator of π under H_0 is given by

$$V(\hat{\pi}_{WAR}) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}. \quad (2.4.9)$$

which can be unbiasedly estimated by

$$\hat{V}(\hat{\pi}_{WAR}) = \frac{\hat{\theta}(1-\hat{\theta})}{n-1}. \quad (2.4.10)$$

and the unbiased estimator of π under H_1 is given by

$$V(\hat{\pi}_{WAR}) = \frac{(p+\pi_y-1)\hat{\theta}_1 - (p-\pi_y)\hat{\theta}_2 - (2\pi_y-1)(1-p)}{(2\pi_y-1)(2p-1)}. \quad (2.4.11)$$

2.5.1 A Mixed Randomized Response Model by Kim and Warde

This model was presented as an effort to defeat the difficulties with the previously constructed models such as Moors (1971), Singh et. al. (2000), Mangat et. al. (1997) models. In this model privacy is better protected and the percent relative efficiency has been compared with the Moors (1971) model. The procedure is as follows

Let a random sample of size n be selected using simple random sampling with replacement. Each respondent in the sample is instructed to answer an innocuous question “I possess the innocuous characteristic Y ”. If the answer to the initial direct question is “yes” then respondent is instructed to go to randomization device R_1 , R_1 consists of two statements, (i) “I belong to sensitive group” and (ii) “I belong to the innocuous group”, with respective probability p_1 and $(1-p_1)$. If the answer to the initial direct question is “no” then respondent is instructed to go to randomization device R_2 , R_2 consists of two statements, (i) “I belong to sensitive group” and (ii) “I do not belong to the sensitive group”, with probability p and $(1-p)$ respectively. In order to offer privacy to the respondents they are not required to tell that which randomizing device they have used. Let n denote the total sample size, n_1 and $n_2 (= n - n_1)$ be the number of respondents using R_1 and R_2 respectively (or in other words reporting “yes” and “no” to the initial direct question respectively).

Denote by θ_1 the probability of “yes” from the respondents using R_1 , then

$$\theta_1 = p_1\pi + (1-p_1)\pi_y, \quad (2.5.1)$$

where π_y is the true proportion of the respondent possessing the innocuous characteristic Y . Note that the respondents coming to R_1 have already reported a “yes” for Y , therefore in R_1 , $\pi_y = 1$, so (2.5.1) reduces to

$$\theta_1 = p_1\pi + (1-p_1). \quad (2.5.2)$$

An unbiased estimator of π , in terms of sample proportion of “yes” responses $\hat{\theta}_1$ is given by

$$\hat{\pi}_{KW1} = \frac{\hat{\theta}_1 - (1-p_1)}{p_1}. \quad (2.5.3)$$

The variance of π_{KW1} is as follows

$$V(\hat{\pi}_{KW1}) = \frac{\theta(1-\theta)}{n_1 p_1^2} = \frac{(1-\pi)[p_1\pi + (1-p_1)]}{n_1 p_1}. \quad (2.5.4)$$

Denote by θ_2 the probability of “yes” from the respondents using R_2 . Then

$$\theta_2 = p\pi + (1-p)(1-\pi).$$

Since all the respondents using R_2 have already reported a “no” for initial direct question, so we have

$$\theta_2 = (2p-1)\pi + (1-p). \quad (2.5.5)$$

An unbiased estimator of π , in terms of sample proportion of “yes” responses $\hat{\theta}_2$, is

$$\hat{\pi}_{KW2} = \frac{\hat{\theta}_2 - (1-p)}{2p-1} \quad (2.5.6)$$

and variance of $\hat{\pi}_{KW2}$ is as follows

$$V(\hat{\pi}_{KW2}) = \frac{X(1-X)}{n_2 p^2} = \frac{\pi(1-\pi)}{n-n_1} + \frac{p(1-p)}{(n-n_1)(2p-1)^2}. \quad (2.5.7)$$

Now a pooled estimator of π in terms of $\hat{\theta}_1$ and $\hat{\theta}_2$ is

$$\hat{\pi}_{KW} = \frac{n_1}{n} \hat{\pi}_a + \frac{n_2}{n} \hat{\pi}_b, \quad (2.5.8)$$

where $\left(0 < \frac{n_1}{n} < 1\right)$.

Now we can see that the two devices and consequently the two estimates of π are independent, so the expected value and variance of $\hat{\pi}_{KW}$ are respectively given below

$$E(\hat{\pi}_{KW}) = \frac{n_1}{n} E(\hat{\pi}_a) + \frac{n_2}{n} E(\hat{\pi}_b) = \frac{n_1}{n} \pi + \frac{n_2}{n} \pi = \left(\frac{n_1+n_2}{n}\right) \pi = \pi \quad (2.5.9)$$

and

$$V(\hat{\pi}_{KW}) = \frac{n_1}{n^2} \left[\frac{(1-\pi)\{p_1\pi + (1-p_1)\}}{p_1} \right] + \frac{n-n_1}{n^2} \left[\pi(1-\pi) + \frac{p(1-p)}{(2p-1)^2} \right]. \quad (2.5.10)$$

To make the variance smaller in (2.5.10), Kim and Warde (2004) suggested allocating more respondents to the first randomization device than the second. This seems awkward in first glance since n_1 and n_2 are random variables and are not directly controllable, however the answer demands sharpness and good judgment. Actually it is possible to control the two sample sizes if the researcher designs the innocuous question in such a way that more respondent say “yes” to it.

In order to offer confidentiality we can make use of the Lanke’s (1975), who derived a value of p which ensures that Simmons’ and Warner’s methods offer equal confidentiality to respondents. The value of p is as under

$$p = \frac{1}{2} + \frac{p_1}{2p_1 + 4(1-p_1)} = \frac{1}{2-p_1}.$$

Substituting this value in (2.5.10), we get

$$V(\hat{\pi}_{KW}) = \frac{\pi(1-\pi)}{n} + \frac{(1-p_1)[\lambda p_1(1-\pi) + (1-\lambda)]}{np_1^2}, \quad (2.5.11)$$

where $\lambda = \frac{n_1}{n}$.

For comparison with Moors (1971), Kim and Warde (2004) computed the percent relative efficiency (PRE) of their method as follows

$$\begin{aligned} PRE &= \frac{Var(\hat{\pi}_M)}{Var(\hat{\pi}_{KW})} \times 100 \\ &= \frac{\left(\frac{1}{np_1^2}\right) \left[\sqrt{y_1(1-y_1)} + (1-p_1)\sqrt{\pi_1(1-\pi_1)} \right]^2}{\left[\frac{\pi(1-\pi)}{n} + \frac{(1-p_1)[\lambda p_1(1-\pi) + (1-\lambda)]}{np_1^2} \right]} \times 100 \\ &= \frac{\left[\sqrt{y_1(1-y_1)} + (1-p_1)\sqrt{\pi_1(1-\pi_1)} \right]^2}{p_1^2 \pi(1-\pi) + (1-p_1)[\lambda p_1(1-\pi) + (1-\lambda)]} \times 100. \end{aligned} \quad (2.5.12)$$

The following points are noticed:

- i. The authors wrote that the PRE given in (2.5.12) would not be affected by the total n . Therefore it suffices the need to compare the PRE for a fixed value of n , hence authors fixed n at 1000.
- ii. The authors concluded that for $\pi_y \geq 0.5$ and for various combinations of p and π , the PRE is greater than 100 except for the case when $\pi = 0.1$.
- iii. The model is always better than Moors (1971) when the proportion of “yes” responses (n_1) is greater than 50 percent.

However we have some other findings about this Kim and Warde (2004) model, that we shall discuss in the next chapter where we present the proposed mixed randomized response model.

2.5.2 The Stratified Mixed Randomized Response Model by Kim and Warde (2004)

As we know that simple random sampling cannot deal with a heterogeneous population, so there is a need to apply stratified random sampling to tackle heterogeneity. Kim and Warde (2004) extended their model to the stratified random sampling scheme. The procedure is as under:

A direct question, “I belong to the innocuous trait group”, is posed to every respondent in each stratum. If the answer to the direct question is “Yes”, then the respondent is directed to choose the randomizing device R_{h1} consisting of two statements (i) “I am a member of the sensitive trait group” and (ii) “I am a member of the innocuous trait group” with preassigned probabilities Q_h and $1-Q_h$, respectively. If a respondent answers “No” to the initial direct question then he is directed to choose R_{h2} consisting of two statements (i) “I am a member of a sensitive trait group” and (ii) “I am not a member of a sensitive trait group” with preassigned probabilities P_h and $1-P_h$, respectively. Suppose m_h is the number of units in the sample from stratum h and n is the total number of units in samples from all strata. Let m_{h1} be the number of people responding “Yes” when responding 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 so that $n = \sum_{h=1}^k m_h = \sum_{h=1}^k (m_{h1} + m_{h2})$. Under the assumption that all respondents are reporting truthfully, and P_h and Q_h ($\neq 0.5$) are set by the researcher.

Now proceeding similar to the SRSWR method we can easily derive an unbiased estimator of π_{sh} as follows,

$$\hat{\pi}_{ah} = \frac{\hat{Y}_h - (1 - Q_h)}{Q_h}, \text{ for } h = 1, 2, \dots, K, \text{ where } K \text{ is the number of strata, } (2.5.13)$$

where \hat{Y}_h is the proportion of “Yes” answers in a sample in stratum h and $\hat{\pi}_{ah}$ is the proportion of respondents with the sensitive trait in a sample from stratum h . Since each \hat{Y}_h is a binomial distribution $B(m_{h1}, Y_h)$, the estimator $\hat{\pi}_{ah}$ is unbiased for π_{sh} with variance as follows

$$\begin{aligned} V(\hat{\pi}_{ah}) &= \frac{Q_h(1 - \pi_{sh})[Q_h\pi_{sh} + (1 - Q_h)]}{(m_{h1})(Q_h^2)} \\ &= \frac{(1 - \pi_{sh})[Q_h\pi_{sh} + (1 - Q_h)]}{(m_{h1})(Q_h)}. \end{aligned} \quad (2.5.14)$$

Similarly using R_{h2} we can derive another unbiased estimator of π_{sh} as follows

$$\hat{\pi}_{bh} = \frac{\hat{X}_h - (1 - P_h)}{P_h}, \text{ for } h = 1, 2, \dots, K, \text{ where } K \text{ is the number of strata,} \quad (2.5.15)$$

\hat{X}_h is the proportion of “Yes” answers in a sample in stratum h and $\hat{\pi}_{bh}$ is the proportion of respondents with the sensitive trait in a sample from stratum h . Since each \hat{X}_h is a binomial distribution $B(m_h, \pi_{sh})$, the estimator $\hat{\pi}_{bh}$ is unbiased for π_{sh} with variance as follows

$$V(\hat{\pi}_{bh}) = \frac{\pi_{sh}(1 - \pi_{sh})}{m_h - m_{h1}} + \frac{(1 - Q_h)}{(m_h - m_{h1})(Q_h^2)}. \quad (2.5.16)$$

Thus an unbiased estimator of π_{sh} , in terms of sample proportions of “Yes” responses \hat{Y}_h and \hat{X}_h , is

$$\hat{\pi}_{msh} = \frac{m_{h1}}{m_h} \hat{\pi}_{ah} + \frac{m_h - m_{h1}}{m_h} \hat{\pi}_{bh}, \text{ for } 0 < \frac{m_{h1}}{m_h} < 1. \quad (2.5.17)$$

with variance as follows

$$V(\hat{\pi}_{msh}) = \frac{\pi_{sh}(1 - \pi_{sh})}{(m_h)} + \frac{(1 - Q_h)[\lambda_h Q_h(1 - \pi_{sh}) + (1 - \lambda_h)]}{(m_h)(Q_h^2)}, \quad (2.5.18)$$

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

Then an unbiased estimator of π_s is shown to be

$$\hat{\pi}_{mS} = \sum_{h=1}^K W_h \hat{\pi}_{msh} = \sum_{h=1}^K W_h \left[\frac{m_{h1}}{m_h} \hat{\pi}_{ah} + \frac{m_h - m_{h1}}{m_h} \hat{\pi}_{bh} \right], \quad (2.5.19)$$

where N is the number of units in the whole population, N_h is the total number of units in the stratum h , and $W_h = (N_h / N)$ for $h = 1, 2, \dots, K$, so that $w = \sum_{h=1}^K W_h = 1$. it

can be shown that the proposed estimator $\hat{\pi}_{mS}$ is unbiased for the population proportion π_s . The variance of the estimator $\hat{\pi}_{mS}$ is

$$V(\hat{\pi}_{mS}) = \sum_{h=1}^K \frac{(W_h)^2}{m_h} \left[\pi_{sh}(1 - \pi_{sh}) + \frac{(1 - Q_h)[\lambda_h Q_h(1 - \pi_{sh}) + (1 - \lambda_h)]}{(m_h)(Q_h^2)} \right]. \quad (2.5.20)$$

Now if we want to optimally allocate the sample size n , we need to know $\lambda_h = (m_{h1} / m_h)$ and π_{sh} . Information on $\lambda_h = (m_{h1} / m_h)$ and π_{sh} is usually unavailable. But if prior information about them is available from past experience then it helps to derive the following optimal allocation formula.

$$\frac{m_h}{n} = \frac{W_h \left[\pi_{sh} (1 - \pi_{sh}) + \left(\frac{1 - Q_h}{(Q_h)^2} \right) [\lambda_h Q_h (1 - \pi_{sh}) + (1 - \lambda_h)] \right]^{1/2}}{\sum_{h=1}^K W_h \left[\pi_{sh} (1 - \pi_{sh}) + \frac{[\lambda_h Q_h (1 - \pi_{sh}) + (1 - \lambda_h)]}{(Q_h)^2} \right]^{1/2}}. \quad (2.5.21)$$

Substituting (2.5.21) in (2.5.20), we get the minimum variance as

$$V(\hat{\pi}_{ms}) = \frac{1}{n} \left[\sum W_h \left\{ \pi_{sh} (1 - \pi_{sh}) + \frac{(1 - Q_h) [\lambda_h Q_h (1 - \pi_{sh}) + (1 - \lambda_h)]}{(m_h)(Q_h)^2} \right\}^{1/2} \right]^2, \quad (2.5.22)$$

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

2.6 Neath's Model

Neath (2004) discussed that even in randomized response atmosphere respondent can hide their true response due to some hidden factors. Neath (2004) presented the measure of information likely to be divulged by the respondent when sensitive study variables are to be studied. Before proceeding further we would like to present the notations to be used in Neath's procedure.

Let x denote an observation from $B(n, p)$,

Then define $\hat{p} = \frac{x}{n}$ and $\pi = P(A)$ or

π = Probabilty of belonging to the sensitive group A,

An unbiased estimator of π is given by

$$\hat{\pi}_A = \frac{\hat{p} + k - 1}{2k - 1}, \quad (2.6.1)$$

where $k = P(I)$

I = Asking, from the respondent, " Do you belong to the sensitive group A",

$p = P(Yes)$,

p = "Yes" response to, from the respondent, " Do you belong to the sensitive group A",

The variance of $\hat{\pi}_A$ as follows

$$V(\hat{\pi}_A) = \frac{p(1-p)}{n(2k-1)^2}. \quad (2.6.2)$$

The choice of k is up to the wish of experimenter however we can assume without loss of generality that $0.5 < k < 1$.

As we know that the most important goals of any randomized response model is to obtain sufficient information on the sensitive study variable(s) while side by side offering maximum privacy protection to the respondents. For the binary sensitive variable response can be either “Yes” or “No”. Now the information likely to be divulged by the respondent, regarding its membership in A , can be characterized by the relative probabilities $P(A/Yes)$ and $P(A/No)$, which are calculated as follows

$$P(A/Yes) = \frac{k\pi}{p} = \frac{k\pi}{k\pi + (1-k)(1-\pi)} \quad (2.6.3)$$

and

$$P(A/No) = \frac{(1-k)\pi}{1-p} = \frac{(1-k)\pi}{(1-k)\pi + k(1-\pi)}. \quad (2.6.4)$$

Now the relative risk of any respondent belonging to group A from to Yes response compared to a “No” response is given by

$$R = \frac{P(A/Yes)}{P(A/No)}. \quad (2.6.5)$$

In (2.6.5), if the relative risk is equal to one, then no information is disclosed on the respondent’s membership in group A , As R diverges away from 1, this information increases. So the motive of any randomized response should be in quantitative terms a small value of the variance in (2.6.2) and a small $|R - 1|$.

Neath (2004) proved that $V(\hat{\pi}_N)$ is a decreasing function of k , if $0.5 < k < 1$. Also R (divulged information) is an increasing function of k , if $0.5 < k < 1$. As the divulged information increases we have an indication that respondents are feeling more privacy protection and hence they are ready to divulge more information. But the goals of maximizing R and minimizing $V(\hat{\pi}_N)$ are in conflict, any choice of k other than 0.5 or 1 represents a compromise.

Neath (2004) also presented the Bayesian approach to measure the divulged information. He proceeded as follows

Assume that the prior distribution of π is $B(a, b)$. Thus,

$$f(\pi) \propto \pi^{a-1} (1-\pi)^{b-1}, \quad 0 < \pi < 1. \quad (2.6.6)$$

Then using transformation techniques, we get the induced prior on p as follows

$$f(p) \propto [p - (1-k)]^{a-1} [k-p]^{b-1}, (1-k) < p < k. \quad (2.6.7)$$

Now our data consists of an observation x from $B(n, p)$, where $p = k\pi + (1-k)(1-\pi)$, so the distribution of π updates to

$$f(\pi/x) \propto [k\pi + (1-k)(1-\pi)]^x [(1-k)\pi + k(1-\pi)]^{n-x} \pi^{a-1} (1-\pi)^{b-1}, 0 < \pi < 1. \quad (2.6.8)$$

Then the posterior distribution of p becomes

$$f(p/x) \propto p^x (1-p)^{n-x} [p - (1-k)]^{a-1} (k-p)^{b-1}, (1-k) < p < k. \quad (2.6.9)$$

Now $100(1-\alpha)\%$ Bayesian confidence interval for p is of the following form

$$(p_L, p_U). \quad (2.6.10)$$

The values of p_L and p_U can be found by solving $F(p_L/x) = 0.1$ and $F(p_U/x) = 0.9$, where $F(p/x)$ is the cumulative distribution function of p .

Now the information divulged by the respondent can also be measured within this Bayesian framework. As $R = \left(\frac{k}{1-k} \right) \left(\frac{1-p}{p} \right)$, then an interval estimate can be easily

derived. Let us denote that interval as (R_L, R_U) given by

$$(R_L, R_U) = \left(\frac{k(1-p_U)}{(1-k)(p_U)}, \frac{k(1-p_L)}{(1-k)(p_L)} \right). \quad (2.6.11)$$

Using (2.6.10) and (2.6.11), we can believe with $100(1-\alpha)\%$ confidence that a person responding “Yes” is between R_L and R_U more likely to possess the sensitive attribute than a person responding “No”.

Regarding privacy protection, according to Neath (2004), more nearer the value of k is to 0.5 more is the privacy protected.

2.7.1 Respondent Jeopardy and Optimal Designs

Many researchers have discussed the jeopardy in randomized response atmosphere. Leysieffer and Warner (1976) discussed that the variance reduction should not be the only criteria to choose the randomizing model, but privacy protection should also be a major concern for the researcher. Authors say that there should be some formal limits on different kinds of jeopardy to which an individual will be exposed by cooperating. Let for example we aim to study a dichotomous finite population in which each person may be classified as earning “equal or more than ten

thousand rupees” or “ less than ten thousand rupees” (monthly). Now when we allow the respondent to randomize his/her response we offer him some comfort that his/her original status will not be exposed, one answer will definitely increase his posterior probability of his/her being in the high income group, while the other answer would increase the posterior probability of his/her being in the low income group. Thus there is a question about respondent jeopardy in these situations.

Therefore some extra effort is required to minimize the variance, minimize the jeopardy and maximize the privacy.

2.7.2 Jeopardizing Responses and the Level of Jeopardy Function

Consider that each respondent either belongs to the group A or not and we want to estimate the proportion of sensitive attribute in population π . Let $P(A/R)$ be the conditional probability that the respondent belongs to A given that his/her reported response is R . R is said to be jeopardizing w.r.t A if

$$P(A/R) > \pi \quad (2.7.1)$$

and jeopardizing w.r.t $B (= A^c)$ if

$$P(B/R) > 1 - \pi. \quad (2.7.2)$$

Or in words, if the posterior probability of a classification increases, given that the observation makes response R , then R is jeopardizing w.r.t that classification.

A more clearer and natural measure for different levels of jeopardy is clearly based on $P(A/R)$ and $P(B/R)$, and are defined as

$$g(R, A) = P(R/A)P(R/B) \quad (2.7.3)$$

and

$$g(R, B) = P(R/B)P(R/A). \quad (2.7.4)$$

Having a look at (2.7.3) and (2.7.4), one may think that g is not a function of π , in fact it is so but $P(A/R)$ and $P(B/R)$ establish the design levels of jeopardy with each response. The maximum of $g(R, A)$ and $g(R, B)$ over the possible R 's in a given design procedure establish the maximal possible jeopardies that can be encountered from different point of view for that design.

For choosing the design among the competent ones the idea is to first approximate the maximal levels in terms of $g(R, A)$ and $g(R, B)$, that are consistent with ethical and

practical cooperation for any respondent. Then among those competent designs the design with least variance should be accepted. Leysieffer and Warner (1976) also discussed the general dichotomous-population-dichotomous-response model that help in comparison among models such as the symmetric model of Warner (1965) and the unrelated question model of Greenberg et. al. (1969), w.r.t. jeopardy consideration.

2.7.3 Bhargava and Singh (2002) Jeopardy Functions

Bhargava and Singh (2002) also discussed the jeopardy functions for Warner (1965) model and did its comparisons with two strategies. The first strategy was as constructed by Mangat and Singh (1990) {see Section (2.3.4) } and the second strategy was as constructed by Mangat (1994).

Mangat (1994) presented a method in which each of the n respondents, selected with SRSWR, is instructed to say “yes” if he/she belongs to the sensitive group A . If he/she does not belongs to the sensitive group A , then the respondent is instructed to use the Warner’s (1965) randomization device. The rest of the procedure is same as was given by Mangat and Singh (1990).

An unbiased estimator of π in this case is therefore given by

$$\hat{\pi}_{MGT} = \frac{\left(\frac{n^*}{n}\right) - (1 - p_3)}{p_3}, \quad (2.7.5)$$

where p_3 the proportion of the sensitive character, is represented in the randomized response device and $\left(\frac{n^*}{n}\right)$ is the observed proportion of “yes” answers, obtained from the n respondents selected with SRSWR.

Then we have the variance of $\hat{\pi}_{MGT}$ as follows

$$V(\hat{\pi}_{MGT}) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(1-p_3)}{p_3}. \quad (2.7.6)$$

From the point of view of variance reduction this model is better than both Warner (1965) and the Mangat and Singh (1990) models. But jeopardy functions should also be given their due rules in efficiency comparison. Now we present the jeopardy functions as developed by the Mangat (1994).

Considering the Warner (1965) model we have the design probabilities as follows:

$$P(y/A) = P(n/B) = p_1, \quad (2.7.7)$$

$$P(n/A) = P(y/B) = 1 - p_1. \quad (2.7.8)$$

From (2.6.16) and (2.6.17), and using W as the index for jeopardy function of Warner's (1965) model, we have

$$g_W(y/A) = \frac{P(y/A)}{P(y/B)} = \frac{p_1}{1 - p_1}, \quad (2.7.9)$$

$$g_W(n/B) = \frac{P(n/B)}{P(n/A)} = \frac{p_1}{1 - p_1}. \quad (2.7.10)$$

Here $p_1 > 0.5$ identifies that "yes" and "no" as jeopardizing for A and B respectively. Let k_1 and k_2 be the maximal allowable values of $g_W(y/A)$ and $g_W(n/B)$. If $k_1 = k_2 = k$, say maximization of $g_W(y/A)$ and $g_W(n/B)$ leads to a design with $\frac{p_1}{(1 - p_1)} = k$, or $p_1 = \frac{k}{k + 1}$.

If, however, $k_1 \neq k_2$ as $g_W(y/A) = g_W(n/B)$, different upper bounds for them cannot be attained simultaneously. In that case if, without loss of generality, $k_1 < k_2$, one should choose the design such that

$$p_1 = \frac{k_1}{k_1 + 1}. \quad (2.7.11)$$

Hence (2.6.20) is the optimal choice for design parameter (p_1) of the Warner's (1965) model. With this choice the minimum variance of the Warner's (1965) estimator is

$$V(\hat{\pi})_{\min} = \frac{\pi(1 - \pi)}{n} + \frac{k_1(k_1 - 1)^{-2}}{n}. \quad (2.7.12)$$

2.7.4 Comparison: Keeping in View the Jeopardy Function of Mangat and Singh (1990) with Warner (1965).

We have the design probabilities, in case of Mangat and Singh (1990) strategy as follows:

$$P(y/A) = P(n/B) = T + (1 - T)p_2, \quad (2.7.13)$$

$$P(n/A) = P(y/B) = (1 - T)(1 - p_2). \quad (2.7.14)$$

Now recalling that Laysieffer and Warner (1976) proposed the natural measure of jeopardy carried by the response R about A and B respectively. These measures are as follows

$$g(R/A) = \frac{P(R/A)}{P(R/B)}; \quad g(R/B) = \frac{1}{g(R/A)}. \quad (2.7.15)$$

These measures are used as jeopardy function in the Bhargava and Singh (2002).

Using (2.7.13), (2.7.14) and (2.7.15), we present the jeopardy function for Mangat and Singh (1990) strategy as given by

$$g_1(y/A) = \frac{T + (1-T)p_2}{(1-T)(1-p_2)}, \quad (2.7.16)$$

$$g_1(n/A) = \frac{T + (1-T)p_2}{(1-T)(1-p_2)}, \quad (2.7.17)$$

and

$$T > \frac{1-2p_2}{2(1-p_2)} \quad (2.7.18)$$

assures that “yes” and “no” are jeopardizing for A and B respectively.

If we proceed in the same manner as in case of Warner’s model, one should design that

$$p_2 = \frac{k_1(1-T) - T}{(1-T)(1+k_1)}, \quad (2.7.19)$$

Therefore the optimal choice of the design parameter (p_2) for Mangat and Singh (1990) strategy, and the resulting minimum variance is $\hat{\pi}_{MST}$ is given by

$$V(\hat{\pi}_{MST})_{\min} = \frac{\pi(1-\pi)}{n} + \frac{k_1(k_1-1)^{-2}}{n}. \quad (2.7.20)$$

Looking at (2.7.12) and (2.7.20), we observe that both are equally efficient at same level of privacy protection.

2.7.5 Comparison: Mangat (1994) Strategy with Warner’s (1965) Model

Let us consider the Mangat (1994) strategy, in which the design probabilities are given by

$$\left. \begin{aligned} P(y/A) &= 1, \\ P(n/A) &= 0, \\ P(y/B) &= 1-p_3, \\ P(n/B) &= p_3. \end{aligned} \right\} \quad (2.7.21)$$

Using (2.7.15) and (2.7.21), the jeopardy functions for Mangat (1994) are given by

$$g_2(y/A) = \frac{1}{1-p_3}, \quad (2.7.22)$$

and

$$g_2(n/B) = \infty, \quad (2.7.23)$$

and $p_3 > 0$ indicates that “yes” and “no” are jeopardizing w.r.t A and B respectively.

Here we note that $g_2(y/A)$ is finite but $g_2(n/B)$ is infinite, which indicates that there is no jeopardy in a “no” answer. Therefore we can take maximal limit for $g_2(y/A)$ as k_1 . It gives the optimal choice of the design parameter (p_3) for Mangat (1994) strategy. That is

$$V(\hat{\pi}_{MGT}) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(k_1-1)^{-1}}{n}. \quad (2.7.24)$$

Now we have the following theorem.

Theorem 2.1. The variance in (2.7.24) is always lesser than the Warner's strategy (1965) when compared at the same level of privacy protection for the respondents, $g_2(n/B)$ being infinite.

Proof:

We know that at equal level of privacy protection the Warner's (1965) strategy can be proved to be less efficient than Mangat (1994) strategy if and only if

$$V(\hat{\pi}_{MGT}) < V(\hat{\pi}). \quad (2.7.25)$$

Using (2.7.12) and (2.7.24), we have

$$(1-\pi)(k_1-1)^{-1} < k_1(k_1-1)^{-2},$$

which reduces to

$$(1-\pi) < \frac{k_1}{k_1-1}. \quad (2.7.26)$$

Since k_1 is greater than unity, so (2.7.26) will always hold.

In the end we restate that, amongst the three strategies Mangat and Singh (1990), Warner's (1965) and Mangat (1994), Mangat and Singh (1990) strategy is found to be equally efficient as Warner's (1965) strategy, but Mangat (1994) is found to be most efficient. It is also reaffirmed that all comparisons are done at the same level of privacy protection.

Chapter 3

A NEW MIXED RANDOMIZED RESPONSE MODEL

3.1 Objectives of the Proposed Model

In this chapter we propose a new mixed randomized response model, with the intention of offering more confidentiality to the respondents, and more efficient estimate of the population proportion of the sensitive characteristic X , when comparing the proposed model with the Kim and Warde (2004) model and implicitly with Moors (1971) model.

Before continuing with the proposed model we bring to light the following frail point/demerit in the Kim and Warde (2004) model.

- I. In model, it is stated that the $PRE = \frac{V(\hat{\pi}_M)}{V(\hat{\pi}_{KW})} \times 100$ is independent of the sample size n . But on changing the values of n one may easily observe that PRE changes and shows that it is dependent on n .

We shall also prove that proposed model is better than the Kim and Warde (2004) model and Moors (1971) model. Moreover, we intend to make the two devices as similar as possible, by replicating the same pair of statements in the two devices with just the difference in the selection probabilities of the statements. By doing so we can promise the respondent that there is no easy way to identify that which device the respondent has used.

3.2.1 The Proposed Model

We now present a new mixed randomized response model that will further curtail down the variance of estimation of the proportion of a sensitive attribute in the population.

Let a random sample of size n be selected using simple random sampling with replacement. Each respondent in the sample is instructed to answer an innocuous question “I possess the innocuous characteristic Y ”.

If the answer to the initial direct question is “yes” then respondent is instructed to go to randomization device R_1 , where R_1 consists of two statements, (i) “I belong to

sensitive group” (ii) “I belong to the innocuous group”, with respective probability p_1 and $(1 - p_1)$. If the answer to the initial direct question is “no” then respondent is instructed to go to randomization device R_2 , where R_2 consists of the same pair of statements as in R_1 but with respective probability p_2 and $(1 - p_2)$. In order to offer privacy to the respondents they are not required to tell that which randomizing device they have used. Let n_1 and n_2 be the number of respondents using R_1 and R_2 respectively, such that $(n_1 + n_2) = n$.

Note that the respondents coming to R_1 have reported a “yes” to the initial direct question therefore $\pi_y = 1$ in R_1 . Denote by X_1 the probability of “yes” from the respondents using R_1 . Then

$$X_1 = p_1\pi + (1 - p_1)\pi_y = p_1\pi + (1 - p_1). \quad (3.2.1)$$

An unbiased estimator for the true proportion of the sensitive trait X is as follows

$$\hat{\pi}_{pro(1)} = \frac{\hat{X}_1 - (1 - p_1)}{p_1}, \quad (3.2.2)$$

where \hat{X}_1 is the sample proportion of “yes” response from the randomizing device R_1 .

The variance of $\hat{\pi}_{pro(1)}$ is given by

$$V(\hat{\pi}_{pro(1)}) = \frac{X_1(1 - X_1)}{n_1 p_1^2}. \quad (3.2.3)$$

From (3.2.1), we have

$$\begin{aligned} X_1(1 - X_1) &= p_1\pi + (1 - p_1) - [p_1\pi + (1 - p_1)]^2 \\ &= p_1(1 - \pi)[p_1\pi + (1 - p_1)]. \end{aligned} \quad (3.2.4)$$

By (3.2.3) and (3.2.4), we get

$$V(\hat{\pi}_{pro(1)}) = \frac{(1 - \pi)[p_1\pi + (1 - p_1)]}{n_1 p_1}. \quad (3.2.5)$$

Note that the respondents using R_2 are reporting a “No” to the initial direct question therefore $\pi_y = 0$ in R_2 . Denote by X_2 the probability of “Yes” from the respondents using R_2 , which is given by

$$X_2 = p_2\pi + (1 - p_2)\pi_v = p_2\pi + \{(1 - p_2) \times 0\} = p_2\pi. \quad (3.2.6)$$

An unbiased estimator for the true proportion of the sensitive trait X is as follows

$$\hat{\pi}_{pro(2)} = \frac{\hat{X}_2 - (1 - p_2)}{p_2}, \quad (3.2.7)$$

where \hat{X}_2 is the sample proportion of “yes” response from the randomizing device R_2 .

The variance of $\hat{\pi}_{pro(2)}$ which is given by

$$V(\hat{\pi}_{pro(2)}) = \frac{X_2(1 - X_2)}{n_2 p_2^2}. \quad (3.2.8)$$

From (3.2.6), we have

$$X_2(1 - X_2) = p_2\pi - (p_2\pi)^2 \quad (3.2.9)$$

Substitution of (3.2.9) in (3.2.8), we get

$$V(\hat{\pi}_{pro(2)}) = \frac{\pi(1 - p_2\pi)}{n_2 p_2^2}. \quad (3.2.10)$$

Now we shall pool the two estimators using weights, and shall optimize the weights to minimize the variance of the weighted estimator $\hat{\pi}_{pro}$.

$$\hat{\pi}_{pro} = d_1 \hat{\pi}_{pro(1)} + d_2 \hat{\pi}_{pro(2)}, \quad d_1 + d_2 = 1. \quad (3.2.11)$$

where d_1 and d_2 are weights such that $d_1 + d_2 = 1$.

The variance of $\hat{\pi}_{pro}$ is given by

$$\begin{aligned} V(\hat{\pi}_{pro}) &= d_1^2 V(\hat{\pi}_{pro(1)}) + d_2^2 V(\hat{\pi}_{pro(2)}) \\ V(\hat{\pi}_{pro}) &= d_1^2 \left(\frac{(1 - \pi)[p_1\pi + (1 - p_1)]}{n_1 p_1} \right) + d_2^2 \left(\frac{\pi(1 - p_2\pi)}{n_2 p_2} \right). \end{aligned} \quad (3.2.12)$$

Minimizing (3.2.12) w.r.t d_i ; $(i = 1, 2)$,

we get the optimum values of d_i as given by

$$d_1 = \frac{v_2}{v_1 + v_2} = d_1^* \text{ (say)}$$

and

$$d_2 = \frac{v_1}{v_1 + v_2} = d_2^* \text{ (say)},$$

where $v_1 = \frac{(1-\pi)[p_1\pi + (1-p_1)]}{n_1p_1}$ and $v_2 = \frac{\pi(1-p_2\pi)}{n_2p_2}$.

Substituting the optimum values of d_1 and d_2 in (3.2.12), we get the minimum variance as follows

$$V(\hat{\pi}_{pro})_{\min} = \frac{v_1v_2}{v_1+v_2} = \frac{\left(\frac{(1-\pi)[p_1\pi + (1-p_1)]}{n_1p_1}\right)\left(\frac{\pi(1-p_2\pi)}{n_2p_2}\right)}{\left(\frac{(1-\pi)[p_1\pi + (1-p_1)]}{n_1p_1}\right) + \left(\frac{\pi(1-p_2\pi)}{n_2p_2}\right)}, \quad (3.2.13)$$

where $p_2 = \frac{1}{2-p_1}$ due to use of the Lanke's (1975), who derived a value of p_2 which ensures that Simmons' and Warner's methods offer equal confidentiality to respondents. The value of p_2 is as under

$$p_2 = \frac{1}{2} + \frac{p_1}{2p_1 + 4(1-p_1)} = \frac{1}{2-p_1}.$$

Now we do the percent relative efficiency (*PRE*) comparison of the proposed estimator $\hat{\pi}_{pro}$ with the moors (1971) estimator, $(\hat{\pi}_M)$ and Kim and Warde (2004) estimator $(\hat{\pi}_{KW})$. Let us define the *PRE* of the $\hat{\pi}_{pro}$ compared to $\hat{\pi}_M$, as follows

$$PRE_{12} = \frac{\left(\frac{1}{np_1^2}\right)\left[\sqrt{y_1(1-y_1)} + (1-p_1)\sqrt{\pi_1(1-\pi_1)}\right]^2}{\left(\frac{(1-\pi)[p_1\pi + (1-p_1)]}{n_1p_1}\right)\left(\frac{\pi(1-p_2\pi)}{n_2p_2}\right)} \times 100, \text{ where } p_2 = \frac{1}{2-p_1}. \quad (3.2.14)$$

Let us define the *PRE* of the $\hat{\pi}_{pro}$ compared to $\hat{\pi}_{KW}$, as follows

$$PRE_{23} = \frac{\left[\frac{\pi(1-\pi)}{n} + \frac{(1-p_1)[\lambda p_1 + (1-\lambda)]}{np_1^2}\right]}{\left(\frac{(1-\pi)[p_1\pi + (1-p_1)]}{n_1p_1}\right)\left(\frac{\pi(1-p_2\pi)}{n_2p_2}\right)} \times 100, \text{ where } p_2 = \frac{1}{2-p_1}. \quad (3.2.15)$$

Now in coming chapters we will numerically examine the properties of percent relative efficiencies given in (3.2.14) and (3.2.15).

3.2.2 Testing the Hypothesis of Truthful Reporting

Now we test the hypothesis of completely truthful reporting versus less than completely truthful reporting. For this purpose, we rewrite the proportion of “yes” from the two devices used in a form incorporating the probability of truthful reporting. Let probability of truthful reporting is denoted by T , where $0 \leq T \leq 1$.

It is assumed that

- i. The probability of truthful reporting is different for the two devices because of the inherent variability in the respondents of two devices. Let T_1 and T_2 be the probability of truthful response in the first and second devices respectively.
- ii. Respondents do not lie about the innocuous trait Y ; they may lie for X (the sensitive trait). However further work may be thought of assuming less than completely reporting even for the innocuous trait.

Note that the respondents coming to R_1 have reported a “yes” to the initial direct question therefore $\pi_y = 1$ in R_1 . Denote by X_1^* the probability of “yes” from the respondents using R_1 , which is given by

$$X_1^* = p_1 \pi T_1 + (1 - p_1) \pi_y = p_1 \pi T_1 + (1 - p_1). \quad (3.2.16)$$

An unbiased estimator for the true proportion of the sensitive trait X is as follows

$$\hat{\pi}_{prop(1)}^* = \frac{\hat{X}_1^* - (1 - p_1)}{p_1 T_1}, \quad (3.2.17)$$

where \hat{X}_1^* is the sample proportion of “yes” response from the randomized response R_1 .

The variance of $\hat{\pi}_{prop(1)}^*$ is given by

$$V(\hat{\pi}_{prop(1)}^*) = \frac{X_1^* (1 - X_1^*)}{n_1 (T_1 p_1)^2}. \quad (3.2.18)$$

From (3.2.14), we have

$$\begin{aligned} X_1^* (1 - X_1^*) &= p_1 \pi T_1 + (1 - p_1) - [p_1 \pi T_1 + (1 - p_1)]^2 \\ &= p_1 (1 - T_1 \pi) [1 - p_1 (1 - T_1 \pi)]. \end{aligned} \quad (3.2.19)$$

Therefore an unbiased estimator of π is given by

$$V(\hat{\pi}_{prop(1)}^*) = \frac{(1-T_1\pi)[1-p_1(1-T_1\pi)]}{n_1p_1T_1^2}. \quad (3.2.20)$$

Note that the respondents using R_2 are reporting a “no” to the initial direct question, therefore $\pi_y = 0$ in R_2 . Denote by X_2^* the probability of “yes” from the respondents using R_2 , which is given by

$$X_2^* = p_2\pi T_2 + (1-p_2)\pi_y = p_2\pi T_2. \quad (3.2.21)$$

An unbiased estimator for the true proportion of the sensitive trait X is as follows

$$\hat{\pi}_{prop(2)}^* = \frac{\hat{X}_2^*}{p_2T_2}. \quad (3.2.22)$$

The variance of $\hat{\pi}_{prop(2)}^*$ is given by

$$V(\hat{\pi}_{prop(2)}^*) = \frac{\pi(1-p_2\pi T_2)}{n_2p_2T_2^2}. \quad (3.2.23)$$

Now we formulate the weighted estimator for π , as given by

$$\hat{\pi}_{prop}^* = d_3 \hat{\pi}_{prop(1)}^* + d_4 \hat{\pi}_{prop(2)}^*, \quad d_3 + d_4 = 1. \quad (3.2.24)$$

Theorem 3.1. The estimator $\hat{\pi}_{prop}^*$ is an unbiased estimator of π .

Proof:

We find the expectation of $\hat{\pi}_{prop}^*$ as follows

$$\begin{aligned} E(\hat{\pi}_{prop}^*) &= d_3 E(\hat{\pi}_{prop(1)}^*) + d_4 E(\hat{\pi}_{prop(2)}^*) \\ &= d_3 E\left(\frac{X_1^* - (1-p_1)}{p_1T_1}\right) + d_4 E\left(\frac{X_2^*}{p_2T_2}\right) \\ &= (d_3 + d_4)\pi \\ E(\hat{\pi}_{prop}^*) &= \pi. \end{aligned} \quad (3.2.25)$$

Now (3.2.25) shows the unbiasedness of $\hat{\pi}_{prop}^*$. However the variance of $\hat{\pi}_{prop}^*$ will

be greater than that of $\hat{\pi}_{pro}$, because the former is built under less than completely truthful reporting and some variation will be introduced due to estimation of T (the probability of true reporting in either device).

Theorem 3.2. The variance of the estimator $\hat{\pi}_{prop}^*$ is given by

$$V(\hat{\pi}_{prop}^*) = \frac{\left(\frac{(1-T_1\pi)[1-p_1(1-T_1\pi)]}{n_1 p_1 T_1^2} \right) \left(\frac{\pi(1-p_2\pi T_2)}{n_2 p_2 T_2^2} \right)}{\left(\frac{(1-T_1\pi)[1-p_1(1-T_1\pi)]}{n_1 p_1 T_1^2} \right) + \left(\frac{\pi(1-p_2\pi T_2)}{n_2 p_2 T_2^2} \right)}.$$

Proof:

We find the variance of $\hat{\pi}_{prop}^*$ as follows

$$V(\hat{\pi}_{prop}^*) = d_3^2 V(\hat{\pi}_{prop(1)}^*) + d_4^2 V(\hat{\pi}_{prop(2)}^*). \quad (3.2.26)$$

$$V(\hat{\pi}_{prop}^*) = d_3^2 \left(\frac{(1-T_1\pi)[1-p_1(1-T_1\pi)]}{n_1 p_1 T_1^2} \right) + d_4^2 \left(\frac{\pi(1-p_2\pi T_2)}{n_2 p_2 T_2^2} \right).$$

Using (3.2.26) we find the optimum values of d_3 and d_4 under the constraint

$$d_3 + d_4 = 1.$$

$$d_3 = \frac{b}{a+b} = d_3^* \text{ (say)}$$

$$d_4 = 1 - d_3^* = \frac{a}{a+b} = d_4^* \text{ (say),}$$

$$\text{where } a = \left(\frac{(1-T_1\pi)[1-p_1(1-T_1\pi)]}{n_1 p_1 T_1^2} \right) \text{ and } b = \left(\frac{\pi(1-p_2\pi T_2)}{n_2 p_2 T_2^2} \right).$$

Substituting the optimum values of w_3 and w_4 in (3.2.26), we get the minimum variance as follows

$$V(\hat{\pi}_{prop}^*)_{\min} = \frac{\left(\frac{(1-T_1\pi)[1-p_1(1-T_1\pi)]}{n_1 p_1 T_1^2} \right) \left(\frac{\pi(1-p_2\pi T_2)}{n_2 p_2 T_2^2} \right)}{\left(\frac{(1-T_1\pi)[1-p_1(1-T_1\pi)]}{n_1 p_1 T_1^2} \right) + \left(\frac{\pi(1-p_2\pi T_2)}{n_2 p_2 T_2^2} \right)}, \text{ where } p_2 = \frac{1}{2-p_1}. \quad (3.2.27)$$

From (3.2.13) and (3.2.27), we can clearly see that

$$V(\hat{\pi}_{prop}^*) \geq V(\hat{\pi}_{pro}) \text{ for } 0 \leq T \leq 1.$$

The two variances will coincide when $T_1 = T_2 = 1$.

Now we will proceed to test the hypothesis of completely truthful reporting versus less than completely truthful reporting.

Note that if $T_1 = T_2 = 1$, (3.2.15) and (3.2.19) become

$$X_1^* = p_1\pi + (1-p_1)\pi_y = p_1\pi + (1-p_1).$$

and

$$X_2^* = p_2\pi + (1-p_2)\pi_y = p_2\pi + (1-p_2) = p_2\pi.$$

We can test that whether the probability of truthful reporting is one or less than one in first randomizing device, by testing

$$H_0 : T_1 = 1 \text{ vs } H_1 : T_1 < 0 \text{ which is equivalent to testing}$$

$$H_0 : X_1^* = \pi p_1 + (1-p_1) \quad \text{vs} \quad H_1 : X_1^* < \pi p_1 + (1-p_1).$$

This hypothesis can be easily tested and the associated critical region is

$$\left[\frac{X_1^*}{\frac{X_1^*(1-X_1^*)}{\sqrt{n_1}}} \right] \geq z_{\alpha/2},$$

where $z_{\alpha/2}$ is the α_{th} quintile point of the standard normal distribution.

Similarly we can test that whether the probability of truthful reporting is one or less than one in second randomizing device by testing

$$H_0 : T_2 = 1 \text{ vs } H_1 : T_2 < 0 \text{ which is equivalent to testing}$$

$$H_0 : X_2^* = \pi p_2 + (1-p_2) \quad \text{vs} \quad H_1 : X_2^* < \pi p_2 + (1-p_2).$$

This hypothesis can be easily tested and the associated critical region is

$$\left[\frac{X_2^*}{\frac{X_2^*(1-X_2^*)}{\sqrt{n_2}}} \right] \geq z_{\alpha/2}.$$

3.2.3 Estimating the True Probability of Truthful Reporting .

Now one may be interested in estimating the true probability of truthful reporting when using the two devices. That is to say one may want to estimate T_1 and T_2 the respective probability of truthful reporting in the population belonging to the two random devices.

From equation (3.2.14) we construct an unbiased estimator of T_1 .

$$\hat{T}_1 = \frac{X_1^* - (1-p_1)}{p_1\pi}, \quad (3.2.28)$$

where \hat{X}_1^* is the sample proportion of “yes” response from the randomizing device R_1 .

The variance of \hat{T}_1 is given by

$$V(\hat{T}_1) = \frac{V(\hat{X}_1^*)}{(p_1\pi)^2} = \frac{X_1^*(1-X_1^*)}{n_1(p_1\pi)^2} \text{ or}$$

$$V(\hat{T}_1) = \frac{(1-T_1\pi)[1-p_1(1-T_1\pi)]}{n_1p_1\pi^2}. \quad (3.2.29)$$

Similarly the estimator for T_2 is obtained from (3.2.19) as follows

$$\hat{T}_2 = \frac{\hat{X}_2^*}{p_2\pi}. \quad (3.2.30)$$

The variance of \hat{T}_2 is given by

$$V(\hat{T}_2) = \frac{V(\hat{X}_2^*)}{(p_2\pi)^2} = \frac{\hat{X}_2^*(1-\hat{X}_2^*)}{n_2(p_2\pi)^2} = \frac{T_2(1-p_2\pi T_2)}{n_2p_2\pi}. \quad (3.2.31)$$

Note that all the variance equations built so far can easily be estimated from the sample statistics.

Chapter 4

A NEW MIXED STRATIFIED RANDOMIZED RESPONSE MODEL

4.1 The Need for Stratification

Up till now, in previous chapters we have been discussing the randomized response strategies in simple random sampling scheme. However many researchers have discussed the randomized response techniques in stratified random sampling, probability proportional to size sampling etc. Therefore we present our proposed model in stratified sampling scheme.

The need for stratification may arise due to heterogeneity in the population with respect to sensitive variable. For example a sensitive variable may be more sensitive for young persons in the population and may be less sensitive for matured persons. That is to say the behaviors of respondent must be tamed by reasonable technique of controlling the unidirectional heterogeneity.

4.2 Stratified Proposed Model

Now we introduce our proposed model in stratified random sampling. The methodology proceeds as follows:

Let a uni-variate heterogeneous population is stratified in h ($h=1,2,\dots,L$) mutually exclusive and collectively exhaustive L strata. Interviewer put a direct question “I am a member of an innocuous trait group Y ” to each respondent in these h stratum. If the respondent in h^{th} stratum answers “yes” to the direct question then he is instructed to go to randomization device R_{h1} consisting of the statements (i) “I belong to sensitive group” and (ii) “I belong to the innocuous group”, with respective probability Q_h and $1 - Q_h$. If the answer to the initial direct question is “no” then respondent is instructed to go to randomization device R_{h2} consisting of the same pair of statements as in R_{h1} but with respective probability P_h and $1 - P_h$. In order to offer privacy to the respondents they are not required to tell that which randomizing device they have used. Let m_h denote the stratum size of h^{th} stratum. Let m_{h1} denote the

number of respondents reporting “yes” to the initial direct question and let m_{h2} denote the number of respondents reporting “no” to the initial direct question also denote by n the total sample size. As we know that the population is dichotomized (and so is the sample, consequently) with respect to the innocuous trait so there are two portions of the sample one of those possessing the innocuous trait and other don’t possessing it, then we have $n = \sum_{h=1}^k m_h = \sum_{j=1}^{N_h} (m_{h1} + m_{h2})$. For convenience we use $L = 2$.

TABLE 1: Stratified Sampling Scheme.

<i>Response to the initial direct question</i>	<i>Stratum I</i>	<i>Stratum II</i>	<i>Total Respondents</i>
Yes	m_{11}	m_{12}	$m_{11} + m_{12}$
No	m_{21}	m_{22}	$m_{21} + m_{22}$
Total respondents	m_1	m_2	n

Assuming completely truthful reporting, the proportion of “yes” responses from the randomizing device R_{h1} will be

$$Y_h = Q_h \pi_h + (1 - Q_h) \pi_{1h}, \tag{4.2.1}$$

where Y_h is the proportion of “yes” in h^{th} stratum, π_h is the proportion of respondents possessing the sensitive trait in h^{th} stratum, π_{1h} is the proportion of respondents with the innocuous trait in h^{th} stratum and Q_h is the probability of selection of the sensitive question in h^{th} stratum.

Note that the respondents using R_{h1} have already reported a “yes” to the initial direct question so for these respondents π_{1h} is equal to one. Therefore we may rewrite (4.2.1) as follows

$$Y_h = Q_h \pi_h + (1 - Q_h). \tag{4.2.2}$$

This yields an unbiased estimator of the population proportion π_h

$$\hat{\pi}_{pro(ah)} = \frac{\hat{Y}_h - (1 - Q_h)}{Q_h}, \text{ for } h = 1, 2, \dots, L, \tag{4.2.3}$$

where \hat{Y}_h is the sample proportion of “yes” responses in h^{th} stratum and $\hat{\pi}_{pro(ah)}$ is the sample proportion of respondents possessing the sensitive attribute in h^{th} stratum.

Utilizing the fact that \hat{Y}_h is distributed as binomial random variable that is $\hat{Y}_h : B(m_{h1}, Y_h)$, it is easy to prove that $\hat{\pi}_{pro(ah)}$ is an unbiased estimator of π_h .

The variance of $\hat{\pi}_{pro(ah)}$ is given by

$$\begin{aligned} V(\hat{\pi}_{pro(ah)}) &= \frac{Q_h(1-\pi_h)[Q_h\pi_h + (1-Q_h)]}{m_{h1}Q_h^2} \\ &= \frac{(1-\pi_h)[Q_h\pi_h + (1-Q_h)]}{m_{h1}Q_h} = V_{H1} \text{ (say)}. \end{aligned} \quad (4.2.4)$$

Assuming completely truthful reporting, the proportion of “yes” responses from the randomizing device R_{h2} will be

$$X_h = P_h\pi_h + (1-P_h)\pi_{ih}, \quad (4.2.5)$$

where X_h is the proportion of “yes” in h^{th} stratum, π_h is the proportion of respondents possessing the sensitive trait in h^{th} stratum, π_{ih} is the proportion of respondents with the innocuous trait in h^{th} stratum, and P_h is the probability of selection of the sensitive question in the h^{th} stratum.

Note that the respondents using R_{h2} have already reported a “yes” to the initial direct question so for these respondents π_{ih} is equal to zero. Therefore we may rewrite (4.2.5) as follows

$$X_h = P_h\pi_h. \quad (4.2.6)$$

This yields an unbiased estimator of the population proportion π_h

$$\hat{\pi}_{pro(bh)} = \frac{\hat{X}_h}{P_h}, \quad \text{for } h = 1, 2, \dots, L. \quad (4.2.7)$$

where \hat{X}_h is the sample proportion of “yes” responses in h^{th} stratum and $\hat{\pi}_{pro(bh)}$ is the sample proportion of respondents possessing the sensitive attribute in h^{th} stratum.

Utilizing the fact that \hat{X}_h is distributed as binomial random variable that is $\hat{X}_h : B(m_{h2}, X_h)$, it is easy to prove that $\hat{\pi}_{pro(bh)}$ is an unbiased estimator of π_h .

The variance of $\hat{\pi}_{pro(bh)}$ is

$$\begin{aligned} V(\hat{\pi}_{pro(bh)}) &= \frac{\pi_h P_h (1 - P_h \pi_h)}{m_{h2} P_h^2} \\ &= \frac{\pi_h (1 - P_h \pi_h)}{m_{h2} P_h}. \end{aligned} \quad (4.2.8)$$

As we know that $m_h = m_{h1} + m_{h2}$, therefore $m_{h2} = m_h - m_{h1}$. So (4.2.8) becomes

$$V(\hat{\pi}_{pro(bh)}) = \frac{\pi_h (1 - P_h \pi_h)}{(m_h - m_{h1}) P_h} = V_{H2} \text{ (say)}. \quad (4.2.9)$$

Now we shall pool the two estimators using weights to formulate an unbiased estimator of π_H , and shall optimize the weights to minimize the variance of the weighted estimator $\hat{\pi}_{pro(h)}$.

$$\hat{\pi}_{pro(h)} = d_{h1} \hat{\pi}_{pro(ah)} + d_{h2} \hat{\pi}_{pro(bh)} \text{ such that } d_{h1} + d_{h2} = 1. \quad (4.2.10)$$

It can be shown that the proposed estimator $\hat{\pi}_{pro(h)}$ is unbiased estimator of π , and its variance is given by

$$\begin{aligned} V(\hat{\pi}_{pro(h)}) &= V \left[t_1 \hat{\pi}_{pro(ah)} + t_2 \hat{\pi}_{pro(bh)} \right] \\ &= t_1^2 V_{H1} + t_2^2 V_{H2}, \\ &= t_1^2 \left(\frac{(1 - \pi_h) [Q_h \pi_h + (1 - Q_h)]}{m_{h1} Q_h} \right) + t_2^2 \left(\frac{\pi_h P_h [1 - P_h \pi_h]}{m_{h2} P_h^2} \right), \end{aligned} \quad (4.2.11)$$

where we let V_{H1} and V_{H2} be the variances of $\hat{\pi}_{pro(ah)}$ and $\hat{\pi}_{pro(bh)}$ respectively and t_1 and t_2 , defined in (4.2.12) and (4.2.13), are weights such that $t_1 + t_2 = 1$. The optimum values of t_1 and t_2 are as follows

$$t_1 = \frac{V_{h2}}{V_{h1} + V_{h2}} = t_1^* \text{ (say)} \quad (4.2.12)$$

and

$$t_2 = 1 - t_1 = \frac{V_{h1}}{V_{h1} + V_{h2}} = t_2^* \text{ (say)}, \quad (4.2.13)$$

Substituting the optimum values of t_1 and t_2 in (4.2.11), we get the minimum variance as follows

$$V(\hat{\pi}_{PRO(h)}) = \frac{V_{h1}V_{h2}}{V_{h1} + V_{h2}} = \frac{\left(\frac{(1-\pi_h)[Q_h\pi_h + (1-Q_h)]}{m_{h1}Q_h} \right) \left(\frac{\pi_h P_h (1-P_h\pi_h)}{m_{h2}P_h^2} \right)}{\left(\frac{(1-\pi_h)[Q_h\pi_h + (1-Q_h)]}{m_{h1}Q_h} \right) + \left(\frac{\pi_h P_h (1-P_h\pi_h)}{m_{h2}P_h^2} \right)}. \quad (4.2.14)$$

Keep in mind that we are concentrating (for convenience) on two stratum, now we present an unbiased estimator of π as follows:

$$\hat{\pi}_{PRO} = W_1 \hat{\pi}_{PRO(1)} + W_2 \hat{\pi}_{PRO(2)}. \quad (4.2.15)$$

Now the variance of $\hat{\pi}_{PRO}$ is as follows

$$V(\hat{\pi}_{PRO}) = (W_1)^2 V[\hat{\pi}_{PRO(1)}] + (W_2)^2 V[\hat{\pi}_{PRO(2)}].$$

Now from (4.2.14) we put the optimized values of the $V(\hat{\pi}_{PRO(1)})$ and $V(\hat{\pi}_{PRO(2)})$

$$V(\hat{\pi}_{PRO}) = (W_1)^2 \frac{V_{11}V_{12}}{V_{11} + V_{12}} + (W_2)^2 \frac{V_{21}V_{22}}{V_{21} + V_{22}}. \quad (4.2.16)$$

Now let

$$a_1 = V_{11} * m_{11},$$

$$a_2 = V_{12} * m_{12},$$

$$a_3 = V_{21} * m_{21},$$

$$a_4 = V_{22} * m_{22}.$$

Now we know that the following equalities hold in our model such that

$$m_{12} = m_1 - m_{11},$$

$$m_{22} = m - m_1 - m_{21},$$

Then substituting all these values in (4.2.16), we get

$$V(\hat{\pi}_{PRO}) = (W_1)^2 \frac{a_1 a_2}{a_1(m_1 - m_{11}) + a_2 m_{11}} + (W_2)^2 \frac{a_3 a_4}{a_3(m - m_1 - m_{22}) + a_4 m_{21}}. \quad (4.2.17)$$

As we know that m_1 and m_2 are unknown, but if some information about them is available than we can optimize the $V(\hat{\pi}_{PRO})$ w.r.t. m_1 ,

we get

$$m_1 = \frac{c_0(a_3n - a_3m_{21} + a_4m_{21}) + m_{11}(a_1 - a_2)}{c_0a_3 + a_1},$$

where

$$c_0 = \frac{w_1 a_1 \sqrt{a_2}}{w_2 a_3 \sqrt{a_4}}.$$

Also the optimum value of m_2 can be obtained after subtraction by assuming a feasible value of m , that is to say we can use the equality $m_{2(opt)} = m - m_{1(opt)}$.

Now we shall compare, numerically, the proposed stratified model with the Kim and Warde (2004) stratified model.

Chapter 5

POPULATION BASED VIRTUAL RESULTS

5.1 Introduction

In this chapter we instigate a virtual analysis of the proposed model over the existing model specifically we show the efficacy of our proposed model, in estimation of the proportion of sensitive qualitative variable over Kim and Warde (2004) and Moors (1971) model.

5.2 Descriptive Statistics of the Data Sets considered, under Truthful Reporting Atmosphere.

We have chosen the following data sets for comparison purposes. We have displayed here the comparison study on two different sample sizes $n = 100, 500$. Consider the value of n_1 and n_2 first. We have considered n_1 from its minimum value to the maximum because there is a restriction in Kim and Warde model (2004) that it is inefficient than the Moors (1971) model if n_1 is smaller. With this point of view, we proceed to check whether our proposed model can counterattack this restriction that is to say, can it work even if this restriction is not valid? In Data Set 1 through Data Set 20, the values of p_1 and p_2 are chosen to be 0.1 for $n = 100$ for the reason that we want to inquire to how much extent such a high privacy protection can raise the efficiency of the model. Then in Data Set 21-40, we study $n = 100$ case for moderate value of p_1 and p_2 choosing $p_1 = 0.5$ and $p_2 = 0.6$ (see appendix for data).

5.3 Efficiency Comparison of the Proposed Model with Competing Models.

Now we present the *PRE* comparisons of the three competitive models considered under completely truthful reporting atmosphere, that is to say we proceed to compare the following three estimators

- i) $\hat{\pi}_M$: the unbiased estimator of π , developed by Moors (1971).
- ii) $\hat{\pi}_{KW}$: the unbiased estimator of π , developed by Kim and Warde (2004).

iii) $\hat{\pi}_{pro}$: the unbiased estimator of π , we propose under completely truthful reporting.

We use the following expression for comparison

$$PRE_{i,j} = \frac{V(i)}{V(j)} \times 100, \quad i = 1, 2, 3 \text{ and } j = 1, 2, 3 \text{ for } i < j,$$

where $V(1) = V(\hat{\pi}_M)$, $V(2) = V(\hat{\pi}_{KW})$ and $V(3) = V(\hat{\pi}_{pro})$.

TABLE 5.14 *PRE of competitive models in Data Set 1- 20*

Data Set	PRE_{12}	PRE_{13}	PRE_{23}
1	83.9800	21857.0	26024.0
2	91.6300	19963.0	21785.0
3	153.870	11543.0	7501.80
4	233.000	7333.80	3147.40
5	916.330	1229.90	134.220
6	100.950	127830	12662.0
7	110.460	117940	10676.0
8	190.000	7396.40	3892.9.0
9	296.870	5197.90	1750.90
10	1610.10	2009.90	124.830
11	84.8500	8960.80	10560.0
12	93.0400	8383.00	9010.00
13	162.820	5815.10	3571.40
14	260.510	4531.10	1739.30
15	2003.90	2669.40	133.200
16	69.3100	7879.00	11367.0
17	75.9100	7309.40	9627.90
18	131.690	4777.80	3627.90
19	208.160	3512.00	1687.10
20	1316.90	1676.60	127.310

TABLE 5.15 *PRE* of competitive models in Data Set 21- 40

Data Set	PRE_{12}	PRE_{13}	PRE_{23}
21	81.640	669.59	820.13
22	85.960	625.33	727.40
23	112.410	428.59	381.25
24	132.850	330.23	248.56
25	180.420	187.60	103.98
26	100.670	448.50	445.51
27	107.140	435.00	406.00
28	150.000	375.00	250.00
29	187.500	345.00	184.00
30	294.110	301.50	102.51
31	86.1800	337.83	391.99
32	92.6400	340.81	367.86
33	138.970	354.04	254.76
34	185.290	360.66	194.64
35	358.630	370.26	103.24
36	85.6000	352.01	411.21
37	91.5300	347.05	379.17
38	132.210	325.02	245.83
39	169.980	314.00	184.72
40	290.210	298.03	102.69

TABLE 5.16 *PRE* of competitive models in Data Set 41- 60

Data Set	PRE_{12}	PRE_{13}	PRE_{23}
41	82.629	5187.6	6278.1
42	89.991	4726.8	5252.5
43	109.99	3786.4	3442.5
44	164.98	2375.8	1440.0
45	490.05	504.35	102.92
46	100.18	3111.0	3105.4
47	109.76	2893.8	2636.6
48	136.36	2450.8	1797.2
49	214.29	1786.2	833.54
50	885.83	904.43	102.10
51	84.323	2207.1	2617.4
52	92.799	2108.8	2272.5
53	116.75	1908.3	1634.5
54	190.48	1607.5	843.90
55	1175.0	1208.4	102.84
56	73.176	2050.8	2802.6
57	80.348	1928.3	2400.0
58	100.43	1678.4	1671.1
59	160.70	1303.4	811.11
60	787.72	805.98	102.32
61	81.331	539.69	663.57
62	88.163	524.57	595.00
63	106.40	493.71	464.00
64	154.29	447.43	290.00
65	382.84	386.02	100.83

From all of the above tables we conclude that $\hat{\pi}_{KW}$ presented by Kim and Warde (2004) is inefficient than Moors (1971) when n_1 is comparatively much smaller than n_2 . Now

when we compare $\hat{\pi}_M$ and $\hat{\pi}_{pra}$, we see that, numerically, $\hat{\pi}_{pra}$ is always efficient than $\hat{\pi}_M$. Also, numerically, $\hat{\pi}_{pro}$ is always efficient than $\hat{\pi}_{KW}$.

5.4 Comparison of the Data Sets considered, under Stratified Random Sampling, Truthful Reporting Atmosphere.

Now we present the *PRE* comparisons of models considered under completely truthful reporting atmosphere, in stratified random sampling atmosphere, that is to say we proceed to compare the following estimators,

- i. $\hat{\pi}_{KW}$: The unbiased estimator of π , developed by Kim and Warde (2004), under stratified random sampling scheme.
- ii. $\hat{\pi}_{PRO}$: The unbiased estimator of π , we propose under completely truthful reporting under stratified random sampling scheme.

It is mentioned here that following notation is used for comparison.

$$PRE_{KW/PRO} = \frac{V(\hat{\pi}_{KW})}{V(\hat{\pi}_{PRO})} \times 100.$$

Table 5.20 *PRE* of competitive models in Data Set 66-73

Data Set	<i>PRE</i> _{KW/PRO}
66	435.6596
67	255.8372
68	170.2432
69	132.2454
70	204.8046
71	156.4602
72	126.9052
73	113.6408

We note that the proposed estimator $\hat{\pi}_{PRO}$, serves better than Kim and Warde (2004) estimator $\hat{\pi}_{KW}$, in stratified random sampling as well. We have used sample size of 200 only because it is understandable that increasing the sample size will make our estimator more and more efficient.

5.5.1 Real Life Survey in Simple Random Sampling Scheme

In order to check the practicality of our proposed model we conducted a real life survey. The main aims of our survey was

- To check the practicality of our proposed model
- To derive an estimate of the proportion of non tax-payers in the middle class portion of the society.
- To compare the population based/ survey based results with the simulated results.

Our target population was all the middle class residents of Islamabad we therefore sampled different residents belonging to middle class. Gender of the respondents was also recorded so as to be utilized in stratified random sampling. All the $n = 500$ respondents were first asked the innocuous question “Were you born in first three months (i.e. January, Feburary, March) of the calendar year?” if the answer was yes they were directed to go to R_1 and if the answer was no then to R_2 .

We selected the sensitive question to be “Are you Non-payer of the income tax?” and the innocuous question to be “Were you born in first three months (i:e Jan, Feb, Mar) of the calendar year?”. Note that the innocuous question is chosen in such a way that π_1 is approximately 0.3, we assumed $\pi = 0.40$.

As we have to use two randomization devices we define R_1 as follows: it consists of 52 playing cards with 26 cards having the sensitive question and 26 having the innocuous question. Similarly define R_2 as a randomization device having 50 cards with 33 cards having the sensitive question and 17 having the innocuous question. We obtained the following results

Table 5.23 Results of the Survey.

Gender	Number of Yes Responses	Number of No Responses
Male	190	73
Female	113	124
Total	303	197

Under simple random sampling scheme, in our survey design $p_1 = 0.5$, $p_2 = 0.66$, $\pi_1 = 0.3$, $\pi = 0.4$. As far as the matter of the value of n_1 (the number of respondents using the first randomization device) is concerned, one may see that respondents use R_1 if they say “yes” to the initial direct question. We have devised the innocuous question in such a way that $\pi_1 = 0.3$, where π_1 is the proportion of respondents possessing the innocuous trait in the population. Since n_1 is a random variable, consequently the expected value of n_1 is given by $E(n_1) = n(\pi_1) = 500(0.3) = 150$, and $E(n_2) = 500 - 150 = 350$.

Now using (3.2.5), we get

$$v_1 = \frac{(1-\pi)[p_1\pi + (1-p_1)]}{n_1 p_1} = \frac{(1-0.4)[(0.5)(0.4) + (1-0.5)]}{(150)(0.5)} = 0.0056.$$

Similarly using (3.2.10) we get

$$v_2 = \frac{\pi(1-p_2\pi)}{n_2 p_2} = \frac{0.4[1-(0.4)(0.66)]}{350(0.66)} = 0.001275.$$

Now using (3.2.11), an unbiased estimate of the proportion of the Non-Tax payers amongst the middle class portion of the society is given by

$$\hat{\pi}_{pro} = w_1 \hat{\pi}_{pro(1)} + w_2 \hat{\pi}_{pro(2)}, \quad w_1 + w_2 = 1,$$

$$\text{where } \hat{\pi}_{pro(1)} = \frac{\hat{X}_1 - (1-p_1)}{p_1} = \frac{[(303)(0.3)/150] - (1-0.5)}{(0.5)} = 0.212,$$

$$\text{and } \hat{\pi}_{pro(2)} = \frac{\hat{X}_2 - (1-p_2)}{p_2} = \frac{[(303)(0.7)/350] - (1-0.66)}{0.66} = 0.4030.$$

Recalling that the optimum weights are given by

$$w_1 = \frac{v_2}{v_1 + v_2} = \frac{0.001275}{(0.0056 + 0.001275)} = 0.1845$$

and

$$w_2 = \frac{v_1}{v_1 + v_2} = \frac{0.0056}{(0.0056 + 0.001275)} = 0.8145.$$

$$\hat{\pi}_{pro} = w_1 \hat{\pi}_{pro(1)} + w_2 \hat{\pi}_{pro(2)} = (0.1845)(0.212) + (0.8145)(0.4030) = 0.3673.$$

Finally using (3.2.13), we get the variance of the proposed estimator, $\hat{\pi}_{pro}$, as follows

$$V(\hat{\pi}_{pro})_{\min} = \frac{v_1 v_2}{v_1 + v_2} = \frac{(0.0056)(0.001275)}{[0.0056 + 0.001275]} = 0.001038. \quad (5.5.1)$$

Now we will compare these results with the simulated results and check for compatibility of the results. When we put $p_1 = 0.5$, $p_2 = 0.66$, $\pi_1 = 0.3$, $\pi = 0.4$, in the simulation formula, we get the value of proposed variance as follows

$$V(\hat{\pi}_{pro})_{\min} = \frac{v_1 v_2}{v_1 + v_2} = 0.0010267. \quad (5.5.2)$$

Comparing (5.5.1) and (5.5.2) we can see that there exists closeness between the population based results and simulations based results.

CONCLUSIONS AND RECOMMENDATIONS

Several procedures and techniques have been used to circumvent the evasive answer bias and to estimate the parameters of the distribution of the sensitive study variables. Warner, S.L. is the Pioneer of the technique the randomized response sampling that deals with circumventing the evasive answer bias.

We have worked on qualitative sensitive variables. We have introduced a new variant in the Kim and Warde (2004) model, and have successfully proposed a new model efficient than the Kim and Warde (2004) model. The results are based on numerical computation and real life survey however unfortunately due to time constraint no real life survey in stratified random sampling scheme has been done. The mathematical derivation of the condition, that ensures the efficiency of the proposed model over the Kim and Warde (2004) model, is rigorous and cumbersome. Therefore such a condition has not yet been derived, but host of simulations and real life survey is backing up.

We have also worked in the less than completely truthful reporting but it has been assumed that the respondents may lie about the sensitive study variable but not the innocuous variable.

This is strongly felt that a real life survey under stratified random sampling scheme may be easily done. Moreover the mathematical condition ensuring the applicability of the proposed model may be derived.

It is also perceived that the situation where respondents may lie about the innocuous variable as well as the sensitive variables may be done.

LITERATURE CITED

1. Adhikari, A.K. , Chaudhuri, A. and Vijayan, K. (1984): Optimum sampling strategies for randomized response trials. *International Statistical Review*, Vol. 52, pp. 115-120.
2. Andrew, A.N. (2004): Quantifying the information from a randomized response. Preprint available at <http://interstat.statjournals.net/YEAR/2004/articles/0402001.pdf>
3. Bhargava, M. and Singh, R. (2002): On the efficiency comparison of certain randomized response strategies, *Metrika*, Vol. 55, pp. 191-197.
4. Boruch, R.F. (1972): Relations among statistical methods for assuring confidentiality of social research data. *Social Science Research*, Vol. 1, pp. 403-414.
5. Chang, H.J and Huang, K.C. (2001): Estimation of proportion and sensitivity of a qualitative character. *Metrika*, Vol. 53, pp 269-280.
6. Folsom, R.E. , Greenberg, B.G. , Horvitz, D.G. and Abernathy, J.R. (1973): The Two Alternate Questions Randomized Response Model for Human Surveys *Journal of the American Statistical Association*, Vol. 68, pp. 525-530.
7. Greenberg, B.G. , Abdel-Latif, A. , Abdul-Ela. , Simmons, W.R. , and Horvitz, D.G. (1969): The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, Vol. 64, pp. 520-539.
8. Gupta, S. and Thornton, B. (2002): Circumventing social desirability response bias in personal interview surveys. *American Journal of Mathematical and Management Sciences*, Vol. 22, pp. 369-383.
9. Horvitz, D.G, Abdul-Ala, A.A. and Simmons, W.R (1967): The unrelated question randomized response model. *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 65-72.

10. Horvitz, D.G. , Greenberg, B.G. and Abernathy, J.R. (1976): Randomized response a data gathering device for sensitive questions. *International Statistical Review*, Vol. 44, pp. 181-196.
11. Horvitz, D.G. , Shah, B.V. and Simmons, W.R. (1967): The unrelated question randomized response model. *Proceedings of the Social Statistics Section, American Statistical Association Society of Statistics*.
12. Huang, K.C. , Lan, C.H. and Kuo, M.P. (2005): Detecting untruthful answering in randomized response sampling. *Quality and Quantity*, Vol. 39, pp. 659-669.
13. Kim, J. and Warde, W.D. (2004): A mixed randomized response model. *Journal of Statistical Planning and Inference*, Vol. 133, pp. 211-221
14. Kim, Jong-Ik. and Flueck, J.A. (1978a): Modifications of the randomized response technique for sampling without replacement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 346-350.
15. Lanke, J. (1975): On the choice of unrelated question in Simmons' version of randomized response. *Journal of the American Statistical Association*, Vol. 70, pp. 80-83.
16. Leysieffer, F.W. and Warner, S.L. (1976): Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, Vol. 71, pp 649-656.
17. Ljungqvist.L. (1993): A Unified Approach to Measures of Privacy in Randomized Response Models: A Utilitarian Perspective. *Journal of the American Statistical Association*, Vol. 88, pp. 97-103.
18. Mahmmod, M. , Singh, S. and Horn, S. (1998): On the confidentiality guaranteed under randomized response sampling: a comparison with several new techniques. *Biometrical Journal*, Vol.40, pp. 237-242.

19. Mangat, N.S. (1994): An improved randomized response strategy. *Journal of Royal Statistical Society, Series B*, Vol. 56, pp. 93-95.
20. 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. *Communications in Statistics Theory and Methods*, Vol. 26 pp. 243–255.
21. Mangat, N.S. and Singh, R. (1990): An alternative randomized response procedure. *Biometrika*, Vol. 77, pp. 439-442.
22. Moors, J.J.A. (1971): Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association*, Vol. 66, pp. 627-629.
23. Papineau, D. (1994): The Virtues of Randomization, *The British Journal for the Philosophy of Science*, Vol. 45, pp. 437-450.
24. Singh, S. , Singh, R. and Mangat, N.S. (2000): Some alternative strategies to Moor's model in randomized response sampling- a survey technique for eliminating evasive answer bias. *Journal of Statistical Planning and Inference*, Vol. 83, pp. 243-255.
25. Tracy, P.E. and Fox, J.E. (1981): The validity of randomized response for sensitive measurements. *American Sociological Review*, Vol. 46, pp.187-200.
26. Warner, S.L. (1965): Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, Vol. 60, pp. 63-69.
27. Zdep, S.M. , Rhodes, I.N. , Schwarz, R. M. and Kilkenny, M. J. (1979):The Validity of the Randomized Response Technique, *The Public Opinion Quarterly*, Vol. 43, pp. 544-549.

Appendix

Note: All data sets are hypothetical and are conceived in accordance with the situation involved.

Descriptive statistics, under Simple Random Sampling using Truthful Reporting Atmosphere.

Table 5.1 Descriptive Statistics for Data set 1- 5

<i>Parameters</i>	<i>Data 1</i>	<i>Data 2</i>	<i>Data 3</i>	<i>Data 4</i>	<i>Data 5</i>
n	100	100	100	100	100
n_1	1	10	50	70	99
n_2	99	90	50	20	1
p_1	0.1	0.1	0.1	0.1	0.1
π	0.2	0.2	0.2	0.2	0.2
π_1	0.3	0.3	0.3	0.3	0.3

Table 5.2 Descriptive Statistics for Data set 5- 10

<i>Parameters</i>	<i>Data 6</i>	<i>Data 7</i>	<i>Data 8</i>	<i>Data 9</i>	<i>Data 10</i>
n	100	100	100	100	100
n_1	1	10	50	70	99
n_2	99	90	50	20	1
p_1	0.1	0.1	0.1	0.1	0.1
π	0.5	0.5	0.5	0.5	0.5
π_1	0.5	0.5	0.5	0.5	0.5

Table 5.3 Descriptive Statistics for Data se10- 15

<i>Parameters</i>	<i>Data 11</i>	<i>Data 12</i>	<i>Data 13</i>	<i>Data 14</i>	<i>Data 15</i>
n	100	100	100	100	100
n_1	1	10	50	70	99
n_2	99	90	50	20	1
p_1	0.1	0.1	0.1	0.1	0.1
π	0.7	0.7	0.7	0.7	0.7
π_1	0.7	0.7	0.7	0.7	0.7

Table 5.4 Descriptive Statistics for Data set 15- 20

<i>Parameters</i>	<i>Data 16</i>	<i>Data 17</i>	<i>Data 18</i>	<i>Data 19</i>	<i>Data 20</i>
n	100	100	100	100	100
n_1	1	10	50	70	99
n_2	99	90	50	20	1
p_1	0.1	0.1	0.1	0.1	0.1
π	0.6	0.6	0.6	0.6	0.6
π_1	0.2	0.2	0.2	0.2	0.2

Table 5.5 Descriptive Statistics for Data set 21- 25

<i>Parameters</i>	<i>Data 21</i>	<i>Data 22</i>	<i>Data 23</i>	<i>Data 24</i>	<i>Data 25</i>
n	100	100	100	100	100
n_1	1	10	50	70	99
n_2	99	90	50	20	1
p_1	0.5	0.5	0.5	0.5	0.5
π	0.2	0.2	0.2	0.2	0.2
π_1	0.3	0.3	0.3	0.3	0.3

Table 5.6 Descriptive Statistics for Data set 26- 30

<i>Parameters</i>	<i>Data 26</i>	<i>Data 27</i>	<i>Data 28</i>	<i>Data 29</i>	<i>Data 30</i>
n	100	100	100	100	100
n_1	1	10	50	70	99
n_2	99	90	50	20	1
p_1	0.5	0.5	0.5	0.5	0.5
π	0.5	0.5	0.5	0.5	0.5
π_1	0.5	0.5	0.5	0.5	0.5

Table 5.7 Descriptive Statistics for Data set 31- 35

<i>Parameters</i>	<i>Data 31</i>	<i>Data 32</i>	<i>Data 33</i>	<i>Data 34</i>	<i>Data 35</i>
n	100	100	100	100	100
n_1	1	10	50	70	99
n_2	99	90	50	20	1
p_1	0.5	0.5	0.5	0.5	0.5
π	0.7	0.7	0.7	0.7	0.7
π_1	0.7	0.7	0.7	0.7	0.7

Table 5.8 Descriptive Statistics for Data set 36- 40

<i>Parameters</i>	<i>Data 36</i>	<i>Data 37</i>	<i>Data 38</i>	<i>Data 39</i>	<i>Data 40</i>
n	100	100	100	100	100
n_1	1	10	50	70	99
n_2	99	90	50	20	1
p_1	0.5	0.5	0.5	0.5	0.5
π	0.6	0.6	0.6	0.6	0.6
π_1	0.2	0.2	0.2	0.2	0.2

Table 5.9 Descriptive Statistics for Data set 41- 45

<i>Parameters</i>	<i>Data 41</i>	<i>Data 42</i>	<i>Data 43</i>	<i>Data 44</i>	<i>Data 45</i>
n	500	500	500	500	500
n_1	1	50	150	300	499
n_2	499	450	350	200	1
p_1	0.2	0.3	0.2	0.2	0.2
π	0.2	0.2	0.2	0.2	0.2
π_1	0.3	0.3	0.3	0.3	0.3

Table 5.10 Descriptive Statistics for Data set 46- 50

<i>Parameters</i>	<i>Data 46</i>	<i>Data 47</i>	<i>Data 48</i>	<i>Data 49</i>	<i>Data 50</i>
n	500	500	500	500	500
n_1	1	50	150	300	499
n_2	499	450	350	200	1
p_1	0.2	0.2	0.2	0.2	0.2
π	0.5	0.5	0.5	0.5	0.5
π_1	0.5	0.5	0.5	0.5	0.5

Table 5.11 Descriptive Statistics for Data set 51- 55

<i>Parameters</i>	<i>Data 51</i>	<i>Data 52</i>	<i>Data 53</i>	<i>Data 54</i>	<i>Data 55</i>
n	500	500	500	500	500
n_1	1	50	150	300	499
n_2	499	450	350	200	1
p_1	0.2	0.2	0.2	0.2	0.2
π	0.7	0.7	0.7	0.7	0.7
π_1	0.7	0.7	0.7	0.7	0.7

Table 5.12 Descriptive Statistics for Data set 55- 60

<i>Parameters</i>	<i>Data 56</i>	<i>Data 57</i>	<i>Data 58</i>	<i>Data 59</i>	<i>Data 60</i>
n	500	500	500	500	500
n_1	1	50	150	300	499
n_2	499	450	350	200	1
p_1	0.2	0.2	0.2	0.2	0.2
π	0.6	0.6	0.6	0.6	0.6
π_1	0.2	0.2	0.2	0.2	0.2

Table 5.13 Descriptive Statistics for Data set 61- 65

<i>Parameters</i>	<i>Data 61</i>	<i>Data 62</i>	<i>Data 63</i>	<i>Data 64</i>	<i>Data 65</i>
n	500	500	500	500	500
n_1	1	50	150	300	499
n_2	499	450	350	200	1
p_1	0.4	0.4	0.4	0.4	0.4
π	0.6	0.6	0.6	0.6	0.6
π_1	0.2	0.2	0.2	0.2	0.2

Descriptive statistics, under Stratified Random Sampling using Truthful Reporting Atmosphere.

Table 5.17 Descriptive Statistics for Data set 66- 69

<i>Parameters</i>	<i>Data 66</i>	<i>Data 67</i>	<i>Data 68</i>	<i>Data 69</i>
n	200	200	200	200
m_1	103	103	103	103
m_2	97	93	90	87
m_{11}	1	50	80	100
m_{21}	1	50	80	86
P_1	0.5	0.5	0.5	0.5
P_2	0.4	0.4	0.4	0.4
t_1	0.6	0.6	0.6	0.6
t_2	0.4	0.4	0.4	0.4
π	0.1	0.1	0.1	0.1
π_{s1}	0.08	0.08	0.08	0.08
π_{s2}	0.13	0.13	0.13	0.13