SYMMETRIC VARIANT TRUTH DETECTION MODEL IN SAMPLE SURVEYS - A RANDOMIZED RESPONSE APPROACH

SIMON MBALA

November 17, 2020

A thesis submitted in fulfillment of the requirements of the degree of doctor of philosophy in mathematical statistics to the School of Mathematics, University of Nairobi

Part I

DECLARATION

This thesis is my original work and has not been presented for examination in any other university.

Signature ____________________________ **Date**__________________________________

Simon Mbala

I80/98927/2015

This thesis has been submitted for examination with our approval as the University Supervisors.

Signature____ **Date** _______________________

Prof. Moses M. Manene

School of Mathematics University of Nairobi

Signature____**Date**______________________________

Prof. Joseph M. Ottieno

School of Mathematics University of Nairobi.

Part II

DEDICATION

This thesis is dedicated to my beloved wife Magdalene Ndanya, for her genuine support and to my children; Victor Muuo, Gideon Mwendwa and Michael Matei. Words are not enough to express my gratitude to you for your support, encouragement and prayers during the course of my studies. May the almighty God bless you abundantly.

Part III

ACKNOWLEDGMENT

I do thank the almighty God for his grace and favor for helping me to complete my studies. Special thanks goes to my Supervisors Prof. Moses Mwangi Manene and Prof. Joseph M. Ottieno for their patience, guidance, commitment and the support they accorded me in writing my thesis. I am also grateful to the lecturers and the staff of the school of Mathematics in the University of Nairobi for their support during my studies. I do also thank Prof. Augustus N. Wali of South Eastern Kenya University for his encouragement to start the Ph.D programme and continuous motivation. I do also appreciate in a special way Mrs Petronila Maria Mulwa, the former principal Muthale girls'secondary school for her encouragement as well as the entire Muthale girls' secondary school fraternity for their moral support during my studies. I also thank Mr Henry M. Kimani, the principal St. Lukes Yatta Boys for his unwavering support during my studies. The success of this work is attributed to the effort of many, however time and space do not allow me to mention you all, may God bless you.

Abstract

In every survey truthfulness is required so as to come up with valid data for decision making. Most surveys use direct questioning to collect data. This method does not yield reliable information when the topic under investigation is sensitive in nature. In such surveys, direct questions are not useful as the respondents will either refuse to answer the survey questions or, even if they do, may give false answers for fear of being known to have the sensitive characteristics. The less privacy a design offers, the more likely respondents cheat by disobeying the instructions thus giving very unreliable information which can lead to wrong decision making. In this study we have formulated a technique which we have called symmetric variant truth detection model. We have also formulated symmetric stratified truth detection model for analyzing stratified data. In this technique, we have used two randomization devices which do not require the respondents to disclose their identity thus increasing their privacy leading to more honest responses. After developing the models, they were validated by the use of data simulation as well as real life application. It was established that the symmetric truth detection models were more efficient compared to the asymmetric truth detection models. This study therefore recommended that researchers on sensitive information to use symmetric truth detection models as opposed to asymmetric truth detection models.

Contents

1 CHAPTER ONE: INTRODUCTION

In this chapter we have discussed the background of the study, notations used in the study, statement of the problem , objectives of the study, significance of the study and research methodology.

1.1 Background information of the study

Sample surveys are conducted by selecting a set of units from a population and recording information or data on the units. The units compose the population and can be individuals, households, institutions or any other element that can be meaningfully thought of as defining a population to be studied. In statistics, survey sampling is a process of selecting a sub-set of elements from a target population to investigate. The term “survey” refers to a formal or official examination of the characteristics or attributes of something or people inorder to ascertain condition and make a decision concerning them.

Surveys are used to collect information that will answer scientific questions. Survey data collection involves different ways of contacting members in a sample once they have been selected. Some of the purposes of sampling is to reduce the cost of investigating the entire target population. Survey samples can be broadly divided into two types; probability samples and non-probability samples. Probability – based samples implement a sampling plan with specified probabilities and allows design based inference about the target population. Some probability sampling methods are as follows; Simple Random Sampling, Stratified Random Sampling, Systematic Random Sampling, Cluster Sampling, Multi-stage and Systematic Sampling The inferences are based on known probability distributions that were specified in the study protocol. In probability samples, each member of the target population has a known and non-zero probability of inclusion in the sample. Survey data can be collected using several methods which include; questionnaires, interview schedules and documentary report among others.

Unlike probability sampling method, non-probability sampling technique uses non-randomized methods to draw the sample. Non-probability sampling method mostly involves judgment. Instead of randomization, participants are selected because they are easy to access. Some non-probability methods of sampling are as follows; Convenience Sampling, Purposive Sampling Quota Sampling and Snowball sampling. Even though in certain cases, non-probability sampling is a useful and convenient method of selecting a sample, the method is appropriate and the only method available in certain cases. One of the major shortcomings of the non-probability sampling is that the findings established through this method lack generalizability. Even though findings obtained through this method apply mostly to the group studied, it may be wrong to extend these findings beyond that particular sample. Through the non-probability method, we can study particular phenomena with a potential to generate valuable insights. The non-probability sample is used to study existing theoretical insights or developing new ones. This method of sampling is considered less expensive, less complicated and easy to apply as compared to its counterpart. However it is not very applicable in this study since we have focused on random sampling leading to generalization which is not the case with non sampling methods.

Self-report is one of the most frequently used data collection techniques in research. However, people do not always tell the truth when asked to answer sensitive questions (Clark and Desharnais, 1998). Socially sensitive questions are thought to be threatening to respondents (Lee and Hong, 1999). In studies such as; examination dishonesty, rape, tax evasion, drug abuse, and prevalence of a certain disease among others, are studied the respondents often react in ways that negatively affect the validity of the data. Such a threat to the validity of the results is the respondents' tendency to give socially desirable answers to avoid social embarrassment and to project a positive self-image (Rasinki,1999). Warner (1965) reasoned that the reluctance of the respondents to reveal sensitive or probably harmful information would diminish if they are convinced that their privacy is guaranteed. Warner, (1965) was the first known Mathematician to introduce the randomized response technique for estimating the proportion of persons bearing a sensitive attribute in a dichotomous population. In Warner's model, with population categories A and A^c a box with two types of cards labeled A and A^c with proportion p and $1-p$ respectively, is used as the randomization device. A respondent draws a card from a box at random and responds 'yes' or 'no' according to whether or not he belongs to the card type he draws from the box. However this method had a problem in that there was low efficiency , large variance and high level of non response leading to data which was not valid.

Horvitz, Shah and Simmons (1967) used RRT in sample surveys and found that the technique produced an estimate of illegitimacy almost as high as the known illegitimacy in the selected sample. The technique has also been used to study abortion cases in North Carolina, USA and was shown to produce an estimate of the proportion of abortions that was in line with previously hypothesized estimates. Greenberg et al. (1969) suggested a RRM asking unrelated question instead of non-sensitive question which was related to sensitive one. Mangat (1994) suggested a forced yes model of three questions forcing respondents to say 'yes' or 'no' as well as sensitive question.The randomized response technique of interviewing on sensitive topics was also used by the National Survey of Family Growth (NSFG) to ascertain the incidence of abortion within a 12-month period, while preserving the individual respondent's complete privacy.

Mathematicians, Economists, psychologists, sociologists, managers, and policy makers have many reasons for asking personal and even intrusive questions. Sensitive questions of interest concern examination dishonesty and drug abuse, tax evasion, employee theft, poaching, regulatory compliance, the integrity of certified public accountants, or participation or interest in deviant or illegal sexual practices, to name just a few (Walter & Preisedovee, 2013). Given the goal of obtaining truthful responses to sensitive queries, it better to use more elaborate techniques that guarantee a respondent's privacy. The overall conclusion of that research is that privacy is not a consideration that people consistently think , indeed, most people have a deep need to share information, including personal information, with others. When respondents are assured of their privacy they can provide truthful information (Chang, et.al, 2004). Bo Yu, et al. (2015) developed a model which considers the estimation of binomial proportions of sensitive attributes in the population of interest in successive sampling on two occasions. In addition, the model was formulated by using rotational cluster sampling when the target population is geographically diverse. Ruenda & Perri (2018) developed a randomization device which combined sensitive research and multiple frame surveys using complex sampling designs. The latest randomized model known to the researcher was developed by Christopher (2019). He developed an efficient randomized response model that can easily be adjusted by selecting certain parameters of the proposed randomized device. All these models, used one randomization devise hence there is a research gap to use more randomization devices which this study has addressed.

1.2 Notations, Terminologies and definitions

1.2.1 Notations

n: Sample size.

n_1 : Sample size for device D_1 :

n_2 : Sample size for device D_2

$p(A|"yes")$:Conditional probability of belonging to the group holding a sensitive attribute A given a "*yes*" answer.

$p(A|"no")$:Conditional probability of belonging to group A given a "*no*" answer.

λ: is the observed proportion of ''*yes*" answers in the sample

$\hat{\lambda}$: The nbiased estimator for λ

α: The probability of having the sensitive attribute.

$\hat{\alpha}$: The unbiased estimator for α.

$1-\alpha$:Probability of interviewees not carrying the sensitive attribute.

A: Represents those who admit having the sensitive attribute.

A^c : Represents those who do not have the sensitive attribute.

p: The probability that a respondent is directed to answer the sensitive question.

$1-p$: The probability that he or she is instructed to answer the non sensitive question.

D_1 : First randomization device.

D_2 : Second randomization device.

p_1: The probability that a respondent is directed to answer the sensitive question by the first randomization device.

$1 - p_1$: The probability that the respondent is instructed to answer the non sensitive question by the first randomization device.

p_2: The probability that the respondent is directed to answer the sensitive question by the second randomization device.

$1 - p_2$: The probability that the respondent is instructed to answer the non sensitive question by the second randomization device.

q: The probability of using the first randomization device.

$1 - q$: The probability of using the second randomization device.

1.2.2 Terminologies and definitions

Privacy protection refers to a situation where the respondent will not be identified.

Randomized response approach refers to research method that allows respondents to respond to sensitive issues while maintaining their confidentiality.

Asymmetry refers to a situation where possible responses (‘‘yes’’ or ‘‘no’’) conveys information on the respondent’s true status.

Symmetry refers to a means of reducing the risk of suspicion where none of the possible responses (‘‘yes’’ or ‘‘no’’) conveys information on the respondent’s true status.

Revealing Refers to exposing true characteristic.

Stigmatizing refers to embarrassing characteristics.

Truth detection refers to the ability of a model to increase the privacy of a respondent leading to giving honest responses.

1.3 Acronyms

RR: Randomized Response.

TDM: Truth Detection Model.

RRT:Randomized response technique.

RRM: Randomized response method

1.3.1 Assumptions of the study

The main assumptions of this study are as follows:-

1. The respondents will answer according to instructions given in the randomization devices.

2. The respondents shall answer truthfully.

1.4 Statement of the problem

Most researchers frequently used direct questioning technique when collecting their research data. However, people do not always tell the truth when asked sensitive questions directly. In collecting sensitive information, two non sampling errors which frequently distort the research findings involve some respondents refusing to answer some questions or deliberately providing incorrect information. Such distortions may result when the respondent is afraid of losing prestige or of becoming embarrassed by offering truthful responses to sensitive questions. The bias produced is sometimes large enough to make the sample estimates seriously misleading. Although topics on personal opinions, controversial issues and intimate behavior are frequently relevant in research, it is very difficult to explore them accurately using traditional survey research methods. In such surveys, direct questions may not be useful as the respondents will either refuse to answer the question or give a socially desired answer which may not be true. To solve this problem, Warner (1965) introduced the randomized response technique (RRT). The rationale of the RRT is that interviewees are more honest when the confidentiality of their responses is guaranteed thus encouraging them to give a more honest response.

A major problem of the RRT models used by Warner (1965) was low efficiency, that is, their large sampling variance. To overcome this weakness, considerable effort has been put into improving the efficiency of RRT models. Chaudhuri (2011) tried to improve it by optimizing design parameters.

Lee, (2015) used asymmetric model as a randomization device for privacy protection where only one randomization device was used. This model had a weakness in that, some participants were suspected to have cheated by denying the sensitive attribute despite being directed by the randomization device to attest to it leading to large variance and increased bias. This is because asymmetric model uses one randomization device and does not provide enough privacy thus encouraging respondents to cheat by saying “no” despite being asked by the randomization device to answer in the affirmative. In asymmetric model a “no” response identifies an interviewee as not holding the sensitive characteristic. Adebola & Johnson (2015) introduced the the use of sub-samples on the non response to improve the efficiency of the variance of the sensitive attribute using one randomization device. However, although he significantly reduced the variance he introduced a serious error in his estimate for the proportion of the “yes” responses by yielding results which were more than 1 which can not be true for the probabilities. In this study we have addressed the problem to false reporting, non response and increased variance in the responses, all attributed to the use of one randomization device used by the earlier researchers. We have extended the use of one randomization device to two randomization devices and compared the relative efficiency of the asymmetric truth detection model and symmetric truth detection model.

1.5 Objectives of the study

In this section we have discussed the general objective and the specific objectives of the study.

1.5.1 Main Objective

The general objective of this study is to formulate a symmetric truth detection model using randomized response approach.

1.5.2 Specific objectives

1. To assess the formulation of asymmetric truth detection models and extend it to symmetric variant truth detection models.
2. To compare the efficiency of symmetric truth detection models with the asymmetric truth detection models.
3. To formulate the symmetric stratified truth detection model and compare it with the asymmetric stratified truth detection model.

1.6 The significance of the study

Non response and false answering are very common in surveys especially when the variables under investigation are sensitive. Reliable data is important in research for decision making. In this study we have developed a mathematical model which will help researchers to get truthful information when investigating sensitive attributes such as criminal cases, diseases like

HIV and Aids, drug abuse, tax evasion and rape among others. The study is also hoped to help the policy makers involved in decision making concerning sensitive information to make informed decisions.

1.7 Research Methodology

This section‘ presents the research methodology that was used in the study. We have shown the models which have been used in this study. These models include the Warner model (1965), Mangat Model (1994), Eichhorn & Hayre Model (2003), Berlev Model (2004), Guerriero model (2007), Lee Model (2015)

1.7.1 The Warner Model

In Warner model, the privacy of the respondent was protected in such a way that research is able to be done without revealing to the interviewer which statement was selected by the randomization device. In this design, respondents are requested to answer ‘‘yes’’ or ‘‘no’’ according to their status on stigmatizing characteristics. In Warner model, p was used as the probability that a respondent is directed to answer the sensitive question and 1-p is the probability that he or she is instructed to answer the non sensitive question. He also used α as the probability of the existence of the sensitive attribute, and 1- α as the probability of non-existence of the sensitive attribute leading to three propositions;

(i) If we let $\hat{\alpha}$ be the unbiased estimator of α, and λ be the observed sample proportion of “yes” response, we can estimate λ as;

$$\hat{\lambda} = \hat{\alpha}(2p - 1) - p + 1 \tag{1.1}$$

(ii) The sample variance of α can be estimated as:

$$Var\left(\hat{\alpha}\right) = \frac{\alpha(1-\alpha)}{n} + \frac{p\left(1-p\right)}{n\left(2p-1\right)^2} \tag{1.2}$$

(iii) The standard error (SE) of $\hat{\alpha}$ is given as;

$$SE((\hat{\alpha})) = \sqrt{\frac{\alpha(1-\alpha)}{n} + \frac{p\left(1-p\right)}{n\left(2p-1\right)^2}} \tag{1.3}$$

However, the Warner model has two weaknesses namely;

the variance and thus standard deviation of the estimator is considerably inflated. Specifically the first term of the equation is the usual variance of a sample proportion. The second term therefore represents the additional sampling error due to the randomizing procedure, thus inflating the variance and some respondents may wonder if there was a mathematical trick that will permit the interviewer to figure out what their true status is.

1.7.2 Mangat Model

Mangat (1994) proposed a strategy in which respondent is instructed to say “yes” if he/she belong to a certain attribute. In this model, some basic demographic questions in sample survey together with questions unrelated to the current study, as well as the sensitive question are included in the questionnaire. One of the four questions is a sensitive one while the other three are not sensitive. This will give rise to four probabilities (p_1, p_2, p_3 and p_4, where the selection probabilities of each question are p_1, p_2, p_3 and p_4, where$\sum\limits_{i=1}^{4} p_i = 1$).

Mangat (1994) used the following notations.

λ = the probability of obtaining "yes" response

α = the sensitive population proportion

α_y = the population proportion of unrelated character

n^* = the number of respondents who say 'yes' among the n persons who were selected.

The probability of obtaining "yes" response was given as;

$$\lambda = \alpha(p_1 - p_2) + (p_2 + p_3\alpha_y + p_4) \tag{1.4}$$

The sensitive population proportion α given as;

$$\alpha = \frac{\lambda - (p_2 + p_3\alpha_y + p_4)}{p_1 - p_2}$$

while the sensitive population proportion character $\hat{\alpha}$ is given as;

$$\hat{\alpha} = \frac{\lambda - (p_2 + p_3\alpha_y + p_4)}{p_1 - p_2}, (p_1 \neq p_2) \tag{1.5}$$

The variance of the sensitive population proportion character $\hat{\alpha}$ is;

$$
\begin{aligned}
V(\hat{\alpha}) = \frac{\alpha(1-\alpha)}{n} + \frac{\alpha(1-p_1-p_2-2p_3\alpha_y-2p_4)}{n(p_1-p_2)} \\
+ \frac{(p_2+p_3\alpha_y+p_4)(1-p_2-p_3\alpha_y-p_4)}{n(p_1-p_2)^2}, \qquad (p_1 \neq p_2) \qquad (1.6)
\end{aligned}
$$

1.7.3 Eichhorn and Hayre Model

Eichhorn and Hayre (2003) extended the model formulated by Mangat (1994) by considering survey models involving a quantitative response variable and proposed a RR design for it. Eichhorn and Hayre (2003) and obtained an unbiased estimate for the expected value of the quantitative response variable of interest to studied some of its immediate properties. Eichhorn and Hayre (2003) estimation procedure considered a RR procedure appropriate for estimating the mean response when the sensitive variable of interest is quantitative in nature. In this model, the interviewees are asked to respond with a coded value composed of their true value for the variable of interest, multiplied by some random number. The interviewer does not know which random number was used by each of the interviewees for coding their responses, but fully knows the underlying distribution which generated the random coding number.

Eichhorn and Hayre (2003) formulated their model by letting X be a random variable (r.v.) denoting the quantitative response variable of interest and Z be a r.v. representing the random number used in the coding mechanism. He assumed that X is independent of Z and Y = ZX be the coded response returned by the interviewee to the sensitive question. Also,

$$\mu_x = E(X)$$

$$\mu_z = E(Z),$$

$$\delta^2 = V(X)$$

and

$$\tau^2 = V(Z)$$

where μ_z and τ^2 are known but μ_x and δ^2 are not known. We also let c_x $= \frac{\delta}{\mu_x}$ and $c_z = \frac{\tau}{\mu_z}$ for the coefficient of variation of X and of Z, respectively. Then, for the coded response,

$Y = XZ,\ E(Y) = E(XZ) = \mu_x.\mu_z$

and

$$V(Y) = E(Y^2) - (E(Y))^2 = \delta^2 + \mu_z^2 + \mu_x^2(1 + c_x^2)\tau_z^2 \qquad (1.7)$$

Based on a random sample $(Y_1, ...Y_n)$ of coded responses of n interviewees, Eichhorn and Hayre (1983) proposed to estimate the unknown mean of the variable of interest,

$\hat{\mu}_x = \frac{\bar{Y}}{\mu_z}$, where $\bar{Y} = \sum \frac{Y_i}{n}$ is the sample mean of the n coded responses. The $\hat{\mu}_x$ is the unbiased estimator of μ_x with variance

$$V(\hat{\mu}_x) = \frac{1}{n}[\delta^2\mu_z^2(1 + \mu_x^2(1 + c_x^2)c_z^2] \tag{1.8}$$

which is larger than that resulting from a simple random sample with direct interviews. The weakness of this model is that the estimate for μ_x variance is uniformly larger than that of $\hat{\mu}_x$ which is likely to give exaggerated results.

1.7.4 Bar-Lev

Bar-Lev (2004) extended the work of Eichhorn and Hayre (2003) by proposing a quantitative RR procedure which generalizes the model for Eichhorn and Hayre (2003) and an estimate for μ_x. This procedure exploits both, the randomizing mechanism used in Warner's original RR model and the quantitative coding scheme in Eichhorn and Hayre (2003). In Bar-Lev (2004) model, the answer given by the respondent k in the the estimation of a mean value μ_x of a sensitive quantitative variable x should not be the true value x_k but a product of z_k such that, letting y_k to be the response then y_k= $z_k.x_k$. The variable z is the randomization response variable, of which the distribution is known. Its expectation and standard deviation are denoted by $\mu_z$and σ_z respectively.

Bar-Lev et al. (2004) also added a second possibility to this randomized response. Let z_k and x_k be as described above. Let p be a design parameter, controlled by the experimenter, which is used for randomizing the interviewees' responses as follows.With probability p the interviewee responds with the true value of the quantitative variable x_k, whereas with probability $1-p$ the interviewee responds with the coded variable $z_k.x_k$. He proposed that the answer y_k of survey unit k is to be given by;

$$y_k = \begin{cases} x_k & with \quad probability \quad p \\ z_k.x_k & with \quad probability \quad 1-p \end{cases}$$

The expectation and variance of the randomly coded response, y_k, are given by;

$$E(y_k) = \mu_x(p + \mu_z(1-p)) \tag{1.9}$$

and

$$V(y_k) = \mu_x^2(1+c_x^2)[p + \mu_z^2(1+c_x^2)(1-p)] - \mu_x^2(p + \mu_z(1-p^2)) \tag{1.10}$$

Hence, the proposed estimate for μ_x based on a random sample of the randomly coded responses, y_1,y_2, ...,y_n is

$\hat{\mu}_x^*= \frac{\overline{y}_k}{p+\mu_z(1-p)}$

where, $\hat{\mu}_x^*$ is the unbiased estimator for μ_x . The weakness of this model is that the design parameter has a role similar to that used in Warner's model and that when $p > 0$ the proposed procedure reduces to that of Eichhorn and Hayre (2003) which does not have many non sensitive questions which reduces the degree of privacy.

1.7.5 Guerriero model (2007)

Guerriero model (2007)modified the randomization device of Bar-Lev (2004) by adding a third possibility to the Bar-Lev process. He did by letting F be a fixed value for the third probability predetermined by the investigator. The randomized response of the survey unit k is therefore;

$$y_k = \begin{cases} x_k & with \quad probability \quad p_1 \\ Z_k x_k & with \quad probability \quad p_2 \\ F & with \quad probability \quad p_3 \end{cases}$$

Where $p(x_k) = p_1$, $p(Z_k x_k) = p_2$ and $p(F) = p_{3,}0 \leq p_i \leq 1,\ i = 1, 2, 3.$su The survey unit k has to answer either truthfully, or in the manner suggested by Bar-Lev (2004), or with a fixed value F predetermined by the agency. The weakness of this model is that the respondents fixed value F might not help in getting the true character of the respondents leading to wrong proportion. There is thus a need to come up with a model with a more truthful response. This can only be possible if the respondents privacy is protected more.

In this model, the expectation of y_k with respect to the randomization is given by

$$E_R(y_k) = p_1.x_k + p_2.x_k\mu_z + p_3F \tag{1.11}$$

$$y_k = x_k b + a$$

with $a \approxeq p_3.F$ and $b \approxeq p_1 +$ $p_2.\mu_z$ hence the term

$x_k = \frac{y_k - a}{b}$, $(b \neq 0)$.

The weakness of this model is that it does not have many non sensitive questions which reduces the degree of privacy. This is because the more non sensitive questions, there is in a questionnaire the higher the protection and the chance of getting more truthful information. To overcome this weakness, TDM presents more sensitive questions by including a prior information about the mean of the study variable which may be used together with sample information.

This study has used all the above models to assess the asymmetric truth detection models since they all used one randomization devise and extend it to two randomization devises.

1.8 Randomized response models

Non response and false answers are very common in survey sampling. This is true especially when the variables asked are sensitive. For such situations, Warner (1965) presented the pioneering work in the field of randomized response questioning designs. Since then, various such techniques with different randomization devices have been proposed by different people. For example Tracy and Mangat (1996), Quatember (2009), Martin (2009) and Heijden and Gils, (1996) among others. All these strategies use a questioning design, which does not enable the data collector to identify the randomly selected question or instruction on which the respondent has given the answer. This reduces the individuals' fear of answering on a sensitive question. The strategy also allows the estimation of the parameter under study because the probability mechanism of the randomization device is known.

Warner (1965) was the first to note that these techniques are also applicable in the field of statistical disclosure control as methods of masking confidential micro-data sets to allow their release for public use. Such micro-data sets may contain variables with sensitive information on an individual. For the randomized response techniques to be applied in this field, either the survey units already perform the randomization of their answers at the survey's design stage or the statistical agency applies the probability mechanism of the technique on the micro-data file after the data collection before its release (Heijden, et.al 1998).

1.9 The Forced Response Model

When the forced response method is used, the respondent is forced by the randomizer to answer the sensitive question (with probability p) truthfully, or to answer ‘‘yes’’ with probability p_{yes} or ‘‘no’’ with probability $1-p-p_{yes}$, independent of the true answer (Boeije, 2002). The unbiased estimate $\hat{\alpha}$ can be computed as;

$$\hat{\alpha} = \frac{\left(\hat{\lambda} - (1-p)\right) p_{yes}}{p} \tag{1.12}$$

where $\hat{\lambda}$ is the observed proportion of "yes" answers in the sample.

The sample variance is given by:-

$$Var\left(\hat{\alpha}\right) = \frac{1}{p^2} \times \frac{\hat{\lambda}\left(1 - \hat{\lambda}\right)}{n} \tag{1.13}$$

1.10 The unrelated questioning technique models

Greenberg et al., (1969) suggested that the respondent should answer one of two questions, however the second question should not be related to the first one but that it should be innocuous by nature.The sensitive characteristic is selected with probability p_1 while the second question unrelated to the sensitive behavior is selected with probability $1 - p_1$. Respondents have to use a randomiser, for instance dice or coins, to decide which of these two questions has to be answered. Although the authors first concern was to help respondents to answer more truthfully, this method has the added statistical advantage of reducing the extra variance added by Warner's method. When the occurrence in the population of the non-sensitive attribute is not known beforehand two independent non-overlapping random samples are needed to compute an unbiased estimate of the sensitive attribute (Fox and Tracy, 1986). For each of these samples the chance that the respondent has to answer the sensitive question has to be different, according to the rules; sample 1 is not equal to sample 2 ; and sample 1 plus sample 2 is not equal to 1. The unbiased estimate of the probability of the sensitive attribute (α_a) in the population can be computed as;

$$\hat{\alpha}_a = \frac{\left[\hat{\lambda}_1 (1 - p_2) - \hat{\lambda}_2 (1 - p_1)\right]}{(p_1 - p_2)} \tag{1.14}$$

where $\hat{\lambda}_1$ is the observed proportion of "yes" answers in sample 1 and $\hat{\lambda}_2$ is the observed proportion of "yes" answers in sample 2 with sample variance

$$Var(\hat{\alpha_a}) = \left[\frac{1}{(p_1 - p_2)^2}\right] \times \left[\frac{\lambda_1 (1 - \lambda_1) (1 - p_2)^2}{n_1} + \frac{\lambda_2 (1 - \lambda_2) (1 - p_1)^2}{n_2}\right] \tag{1.15}$$

2 CHAPTER TWO: LITERATURE REVIEW

In this chapter we have reviewed the work done by other authors on randomized response in sample survey. The problem at hand is to increase the privacy of respondents in investigating sensitive information and reduce the variance in responses.

The randomized response technique is a useful method for collecting data on variables which are considered sensitive, incriminating or stigmatizing for the respondents. Examples of such situations are common in socio-economic surveys, for instance, we may need to collect data on tax evasion, alcohol addiction, illegal drug use, criminal behaviour or past criminal convictions (Lee, 2012). In a randomized response model, the respondents use a randomization device to generate a randomized response and the parameter under study can be estimated from these responses. Thus, the respondent is not required to disclose his true response and it is expected that this will lead to better participation in the survey on sensitive issues.Reliability of data is compromised when sensitive topics on embarrassing or illegal acts are investigated using direct method of data collection in sample survey. The impact of the response distortion on the survey or test results is of a major concern to researchers. Surveys on human population have established the fact that the direct question about sensitive characters often result in either refusal to respond or falsification of the answer (Sidhu et al, 2009). However, obtaining valid and reliable information is a prerequisite for obtaining meaningful results. Hence, there is need to ensure confidentiality of respondents which will in-turn lead to more reliable information. Warner (1965) developed an interviewing procedure designed to reduce or eliminate this bias and called it Randomized Response Technique (RRT). Warner (1965) is the first mathematician known to have developed an interviewing procedure designed to reduce errors caused by non response and false answers when collecting data on sensitive attributes. This technique is called the randomized response technique. This is because the respondent answers one of several questions selected at random and the interviewer is given an answer but is unaware of the question which is being answered by the respondent.

Lanke (1976) initiated the study of efficiency versus privacy protection in randomized response surveys where the population is divided into two complementary sensitive groups, A and A^c. The objective was to estimate the proportions of persons belonging to these groups. Lanke (1976) suggested measures of jeopardy based on the posterior probabilities of a respondent belonging to groups A and A^c given his randomized response. Since then, this dichotomous case has been widely studied. Loyns (1976) extended the jeopardy measure of Lanke (1976) to polychotomous populations. Since then, many researchers have contributed to this area. Anderson (1977) studied the case of continuous sensitive Variables and considered the amount of information provided by randomized responses. For ensuring more privacy he recommended that the expectation of the conditional variance of the sensitive attribute (X) given the randomized response be made as large as possible.

Ljungqvist (1993) gave a unified and utilitarian approach to measures of privacy for the dichotomous case. Mangat (1994) proposed a strategy in which a respondent is instructed to say “yes” if he/she belong to an attribute A, if not, he/she is required to use the Warner randomized device consisting of two statements. Since then many researchers have suggested and extended various models based on Warner model. Greenberg et.al (1969) proposed a questioning technique where the researcher asks unrelated question instead of non sensitive question related to the sensitive one. Nayak (1994) combined Warners (1965) model with Greenberg et.al (1969) and formed a combined RRM which had three questions. Margat et. al (1995) augmented the statements of ordering to say ’no’ to the Warner (1965) model. Eichhorn and Hayre (2003) extended the model developed by Mangat (1994) by developing survey models which allowed responses with a coded value composed of their true value for the variable of interest, multiplied by some random number. In this model, the interviewer does not know which random number was used by each of the interviewees for coding their responses, but fully knows the underlying distribution which generated the random coding number.

Nayak (1994) combined Warner’s RRT with Greenberg et al.’s RRM and formed a combined RRM which had three questions. Bhargava-Singh (2000) switched Mangat et al.’s model by exchanging the statement of ordering an answer ‘no’ instead of answering ‘yes’ . Chang et al. (2004) suggested a combined RRM by holding together Mangat et al.’s model and Bhargava-Singh’ model. In 2005, a meta-analysis on 42 comparative studies by Allyson and

Jon showed that RRTs resulted in more valid population estimates than direct question-answer techniques. This positive effect on the validity of the results was found both when the estimates of RRTs were compared to known population estimates and when the results of RRTs were compared to other data collection methods. It also appeared that the results of the RRTs became more valid when the topic under investigation became more sensitive. Nayak(1994) proposed a measure of jeopardy for surveys from dichotomous populations and developed an approach for comparing the available randomization procedures. These results are all based on samples drawn by simple random sampling with replacement method. However they were not tested using other sampling methods.

Chaudhuri (2011) used a randomization device to determine whether respondents are asked to simply provide a pre-specified answer ("yes") with probability p_{yes}, or whether they are asked to answer the sensitive question honestly with probability $1 - p_{yes}$. Because the interviewer is unaware of the outcome of the randomization device, the randomization ensures that no individual interviewee can be identified as holding the sensitive attribute on the basis of his or her answer. This is because a “yes” answer is now no longer the unambiguous result of truthful answering; it may simply be the outcome of the randomization procedure. Because the probability distribution of the randomization device is known, straightforward probability calculations allow researchers to estimate the proportion of “yes” answers that have not been prompted by the randomization device. The prevalence of the sensitive attribute may thus be estimated at group level, while simultaneously protecting the confidentiality of individual answers. The technique therefore encourages more honest response.

All the references given above are for sensitive variables which are categorical or qualitative in nature. However, in randomized response surveys it is quite common to have situations where the study variable X is quantitative, for example in studies on the number of criminal convictions of a person, the number of induced abortions, the number of months spent in a correction centre, the amount of undisclosed income, and so on. Quatember (2013) suggested a new type of model by adding unrelated question to Chang et al.'s model and we call it Quatember's randomized response model (QRRM). Quatember's (2013) reviewed various dichotomous RRMs with more than three questions and extended QRRM to incorporate the unknown unrelated population proportion and suggested a two sample QRRM. He also proposed a stratified QRRM for stratified population and developed it into stratified two sample QRRM covering unknown proportion of unrelated attribute. He also investigated proportional and optimum allocations as allocation methods under the stratification assumption. In all the above methods cheating was still witnessed in some cases. Two constructs have been proposed to describe potential response hazards, that is, characteristics of RRT designs that can make both guilty and innocent respondents cheat: respondent jeopardy and risk of suspicion. Respondent jeopardy refers to the risk of guilty respondents to be identified as such when truthfully admitting the sensitive attribute. Respondent cheating may be reduced by choosing randomization probabilities close to 0.50, which however reduces efficiency by enlarging the variance of parameter estimates (Quatember's, 2013). Innocent interviewees run a risk of suspicion when being prompted by the randomization device to answer sensitive questions in the affirmative. They tend to feel uncomfortable under such circumstances, because their affirmative answer now seemingly associates them with an undesirable attribute. They may therefore be tempted to play safe by denying the critical attribute in spite of being told otherwise by the randomization procedure (Quatember's, 2013). There exists a gap in all the randomized responses covered so far in this literature in that, both truthfulness or cheating does not stem from the same place hence the respondents can easily play safe by denying sensitive attribute which they have, besides being forced by a randomized device to attest to it. There is therefore a need to come up with a randomized device which does not allow the respondents to cheat.

Adepetum and Adebora (2014) in his article focuses on studying the issue of privacy protection when the variable under study is quantitative and discrete. He proposed the use of a randomization device and the associated estimation method. He then considered two separate cases, one where all values of X are sensitive and another where not all values of X are sensitive. For each of these cases, he proposed a measure for protecting the privacy of the respondents. Adepetum and Adebora (2014) finally showed how one can choose the randomization device parameter in each case, so as to guarantee a certain pre-specified level of respondent protection and then maximize the efficiency of estimating the parameter of interest under this constraint. His study also covers qualitative sensitive variables, that is, cases where the population is dichotomous or polychotomous, and allows the estimate of the proportions of individuals belonging to each category.

Dawes and Moore (2015), have modified the Warner technique and introduced the forced response variant of the RRT. In the Dawes model, first all the interviewees are confronted with the sensitive question, after which a randomization device is used. The respondents were then asked to simply provide a pre-specified answer ("yes") with probability p_{yes}, or to answer the sensitive question honestly (with probability $(1 - p_{yes})$. Because the interviewer is unaware of the outcome of the randomization device, the randomization ensures that no individual interviewee can be identified as holding the sensitive attribute on the basis of his or her answer. This is because a “yes” answer is now no longer the unambiguous result of truthful answering; it may simply be the outcome of the randomization procedure. Because the probability distribution of the randomization device is known, straightforward probability calculations allow the researcher to estimate the proportion of “yes” answers that have not been prompted by the randomization device.

In one study for example (Jarman, 1997), participants were asked to answer an Internet-based survey that included sensitive personal questions. Half of the participants were first asked to complete when and why RRTs fail using a separate questionnaire measuring their Internet privacy concerns.

The questionnaire increased the salience of privacy issues, and decreased participants' self-disclosure: relative to the group of participants who had not been asked to discuss their privacy concerns, participants who had been primed with the questionnaire ended up answering fewer personal questions. These studies and others show that people's concern about privacy can be activated by environmental cues that may bear little, or sometimes even a negative, relationship to objective dangers associated with information sharing.

Instead of assuring privacy, it has been argued that truthful information sharing can be motivated by introducing stochastic noise to the communication channel (Warner 1965). For example, if messages are lost with probability p, the non-arrival of a message cannot be entirely attributed to an unwillingness to send a message, which increases senders' willingness to share information. Likewise, if respondents are instructed to answer a sensitive question truthfully only with probability p – a procedure called the randomized response technique (RRT) – affirmative responses cannot be interpreted on the individual level, thus increasing respondents' willingness to provide sensitive information. Although RRTs take many different forms many authors on randomized response add noise to individual responses, in theory making it easier for individuals to admit to sensitive behaviors, thoughts, and feelings Perri, et.al (2015). For example, in the coin flip technique one of the most common forms of the RRT – the interviewee is asked a sensitive question with response options "yes" and "no." Prior to answering the question, the interviewee flips a coin and answers the question based on the outcome of the coin flip. If he flips ‘heads,’ he is instructed to respond ‘yes,’ regardless of whether he has actually engaged in the given behavior; if he flips ‘tails,’ he is instructed to answer the question truthfully. The interviewer, who cannot see the outcome of the coin flip, cannot tell whether a given ‘yes’ response denotes an affirmative admission or a coin flip that has come up heads or both. By correcting for the (known) probability of answering the focal question (i.e. in the coin flip technique, flipping tails) however, the interviewer can deduce the population-wide prevalence of the behavior. In principle therefore, the RRT can be used to estimate with greater accuracy the prevalence of behaviors that people are uncomfortable disclosing.

Ljungqvist (1993) provided a formalization of the RRT. Utility-maximizing respondents face a tradeoff between lying aversion – they prefer to tell the truth – and stigmatization aversion – they prefer not to be associated with the behavior/information in question. Whether a respondent answers truthfully or not thus depends on both conditional probabilities of being perceived as belonging to the sensitive group or not. Based on Ljungqvist’s model, Blume et al. (2013) developed a game-theoretic formulation of the RRT in which respondents face a tradeoff between lying and stigmatization aversion, and respondents’ payoffs dependent on the interviewer’s beliefs. The model allows for specifying the parameter space for lying and stigmatization aversion for which RRT will induce more truth-telling than direct questioning (DQ), and vice versa, when DQ will induce more truth-telling than RRT. In line with this, in this study we demonstrate that the RRT can yield more valid prevalence estimates than those obtained by DQ, and in some cases, even impossible (negative) prevalence estimates. These effects, however, occur for reasons beyond the specific parameter values for lying and stigmatization aversion. We show that the RRT can perform better than DQ because the RRT makes respondents concerned that innocuous responses will not be interpreted as admissions; because only one response (denial) has an unambiguous interpretation, it leads them to give that response.

Zawar et.al. (2010) developed a Bayesian estimation method for population proportions of a sensitive characteristics which adopts a simple Beta distribution as a quantification of prior information using simple random sampling with replacement. Bo Yu, et al. (2015) developed a model which considers the estimation of binomial proportions of sensitive attributes in the population of interest in successive sampling on two occasions. In addition, the model was formulated by using rotational cluster sampling when the target population is geographically diverse. Ruenda & Perri (2018) developed a randomization device which combined sensitive research and multiple frame surveys using complex sampling designs. The latest randomized model known to the researcher was developed by Christopher (2019). He developed an efficient randomized response model that can easily be adjusted by selecting certain parameters of the proposed randomized device.

All the models reviewed in this literature review used one randomization device and were therefore characterized by large variances as well as high non response rate. In this study we have used two randomization devices as opposed to the earlier randomization methods.

3 CHAPTER THREE: TRUTH DETECTION MODELS

3.1 Introduction

In this chapter we have discussed the asymmetric truth detection models and compared their relative efficiency.

3.1.1 Asymmetric Truth Detection Models

Randomized response survey techniques were developed to permit estimation of frequencies of single stigmatizing characteristics by enabling a respondent to protect his self-image without reporting falsely. The earlier models by Warner(1965), Bar-Lev(2004), Eichhorn & Hayre (2013) and Gjestvang (2007) models, all used one randomization device hence are referred to as asymmetric randomized response models (Martin, 2009). In asymmetric technique, a single randomization device such as a coin, a card or a spinner is used. In these models, a "no" response identifies an interviewee as not holding the sensitive characteristic. Asymmetric models, therefore encourage respondents to cheat by saying "no" despite being asked by the randomization device to answer in the affirmative. In this model, depending on the outcome of the randomization process, respondents are either asked to provide the specified answers "yes" (with probability p or "no" with probability $1 - p$ or to answer the sensitive question honestly (with probability 1–$p_{yes} - p_{no}$).

Mangat (1994) proposed a strategy in which respondent is instructed to say “yes” if he/she has an attribute A, if not, he/she is required to use the Warner randomized device consisting of two statements:

i. I have attribute A with probability p.

ii. I do not have attribute A with probability 1 - p

Suppose we consider the task of estimating the proportion of the population who have some particular characteristic, call it A which is stigmatizing. Some people will refuse to answer such a question, and some of those who answer it do not answer truthfully. If we write the probability that a "yes" response is given to a direct question about trait A as the sum of the probabilities that a person will respond truthfully when trait A is present and that a person will respond but respond falsely when trait A is absent. The summation must be weighted by a factor we call W= $\frac{n}{n'}$ so that;

Pr(yes) + Pr(no) =1 and n' is the expected sample size that results from non-response, defined for the direct question method.

For example, if we want to model the extent of a sensitive attribute, using one randomization device the researchers would ask a less sensitive question and then combine it with the sensitive question. Each participant is given a questions that requires dichotomous answers such as yes or no. An example of one such question given might be;

"Have you ever engaged in examination dishonesty?" Suppose also that all the questions would be framed so that answering "yes" is admitting to engaging in a sensitive behavior. In addition to the list of questions, each participant would be asked his or her admission number and instructions on how to use the number in answering the questions is given. The participants would be told to answer that question according to their admission number and the following rule;

“If your admission number is even, then answer the question truthfully, if it is odd, then ignore the question altogether and just say ’yes’ no matter what you would have answered to the question." Any observed “yes” response could mean either that the respondent had engaged in examination dishonesty behavior or simply that he had an odd admission number. Even if a participant’s answer is known, his or her actual behavior could not be deduced from the answer, thus privacy would be assured. In this case, the admission number was used as a randomization device which is asymmetric. According to Mangat (1994) an investigator could determine the proportion of the sample that engages in any behavior using an equation derived as follows.

Let α be the proportion of the population that would privately admit to having engaged in the sensitive behavior and λ be the proportion of affirmative responses given. The respondents who engaged in the sensitive behavior will have answered "yes" regardless of the outcome of their admission number and if the numbers are well distributed, half of the participants who have not engaged in the behavior, 0.5(1 - α), some respondents will have answered "yes" because their admission number is odd. Therefore,

$\lambda = \alpha + 0.5(1 - \alpha)$

and

$$\alpha = 2\lambda - 1 \tag{3.1}$$

which is the equation for proportion of the population that would privately admit to having engaged in the sensitive behavior.

Using this equation, for example if λ= 0.57, thenα = 0.14 which means that of all the respondents, 50% answered "Yes" because their admission number was odd and 7% did so because their admission number was even and they answered "yes" truthfully to the sensitive question.

Warner (1965) introduced the randomized response technique for estimating the proportion of persons bearing a sensitive attribute in a dichotomous population using one randomization device like a coin, a card or a spinner. In asymmetric model only one randomization device is used, thus with population categories A and A^c, a box with two types of cards labeled A and A^c in proportion p and 1$-p$ respectively is used as the randomization device. Each sampled respondent has to select a card at random from a pack of cards. The pack consists of two types of cards with known proportions and cards are identical in appearance. Card type 1, with proportion p contains the question "Do you belong to the group A?" while card type 2 with proportion 1$-p$ bears the question "Do you belong to the group $\overline{A}$?" where A is the group with the sensitive while $\overline{A}$ is the group without sensitive attribute. The respondent will supply a truthful answer "Yes" or "No" for the question mentioned in the selected card. The experiment is performed in the absence of the interviewer and hence the privacy of the respondent is maintained because the interviewer will not know which of the two questions the respondent has answered.This technique which uses only one randomization device in this case one card park is what we have called asymmetric randomization model (Martin, 2009). In these cards group A represents those who have the sensitive attribute and $\overline{A}$ represents those who do not have the sensitive attribute.

The respondents then have the option'yes' or 'no' according to whether or not they belong to the group of sensitive attribute. Let λ be the observed proportion of ''yes" answers in the sample, p be the probability that a respondent is directed to answer the sensitive question and $1-p$ be the probability that he or she is instructed to answer the non sensitive question. Let α be the probability of the existence of the sensitive attribute and 1- α be the probability of non existence of the sensitive attribute. The response will lead to a Bernoulli distribution.

Proposition 3.1

The estimator for the sensitive attribute α is given by;

$$\hat{\alpha} = \frac{\hat{\lambda} + (p-1)}{2p-1} \tag{3.2}$$

Proof

The total proportion of "yes" irrespective of the question of affirmative response,λ can be expressed in terms of α in the following way;

$$\lambda = p\alpha + (1-p)(1-\alpha)$$

which can be expanded as;

$$\lambda = \alpha(2p-1) + 1 - p \tag{3.3}$$

Making α the subject of the formula we get;

$$\alpha = \frac{\lambda + (p-1)}{2p-1} \tag{3.4}$$

Since it is not possible to calculate α and λ, then α can be estimated as;

$$\hat{\alpha} = \frac{\hat{\lambda} + (p-1)}{2p-1} \tag{3.5}$$

This completes the proof.

Proposition 3.2

The variance of the estimator of α is given by;

$$V(\hat{\alpha}) == \frac{\alpha(1-\alpha)(}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{3.6}$$

Proof

Using equation (3.5) above, the variance of $\hat{\alpha}$ is;

$$V(\hat{\alpha}) = V\left(\frac{\hat{\lambda} - (1-p)}{2p-1}\right)$$

which can simplify to;

$$V(\hat{\alpha}) = V\left(\frac{\hat{\lambda}}{2p-1}\right) - V\left(\frac{(1-p)}{2p-1}\right)$$

but, variance of a constant = 0 therefore,

$$V\left(\frac{(1-p)}{2p-1}\right) = 0,$$

we therefore have;

$$V(\hat{\alpha}) = V\left(\frac{\hat{\lambda}}{2p-1}\right) \tag{3.7}$$

But

$$V(\hat{\lambda}) = \frac{\lambda(1-\lambda)}{n} \tag{3.8}$$

Substituting for ;

$$V(\hat{\lambda})$$

Using equation (3.3), we get;

$$1 - \lambda = \alpha(1-2p) + p \tag{3.9}$$

Substituting for λ in equation 3.9 we get;

$$V(\hat{\alpha}) = \left(\frac{1}{n(2p-1)^2}\right)[\alpha(2p-1)+1-p][\alpha(1-2p)+p)]$$

$$= \frac{[2p\alpha^2 - 4p^2\alpha^2 - \alpha^2 + 2p\alpha^2 + 2p^2\alpha - p\alpha + \alpha - 2p\alpha - p\alpha + 2p^2\alpha + p - p^2]}{n(2p-1)^2}$$

Which simplifies to;

$$V(\hat{\alpha}) = \frac{[4p\alpha^2 + 4p^2\alpha - 4p^2\alpha^2 - 4p\alpha - \alpha^2 + \alpha + p - p^2]}{n(2p-1)^2}$$

$$= \frac{[4p^2\alpha - 4p\alpha + \alpha - 4p^2\alpha^2 - 4\alpha^2 p + \alpha^2 + p - p^2]}{n(2p-1)^2}$$

$$= \frac{p(1-p)}{n(2p-1)^2} + \frac{[\alpha(4p^2-4p+1) - \alpha^2(4p^2-4p+1]}{n(2p-1)^2}$$

Which reduces to;

$$V(\hat{\alpha}) = \frac{p(1-p)}{n(2p-1)^2} + \frac{[(\alpha-\alpha^2)(2p-1)^2}{n(2p-1)^2}$$

The variance of $\hat{\alpha}$ is therefore given by,

$$V(\hat{\alpha}) = \frac{\alpha(1-\alpha)}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{3.10}$$

This completes the proof.

The standard error $\hat{\alpha}$ is consequently given as;

$$SE(\hat{\alpha}) = \sqrt{\left(\frac{p(1-p)}{n(2p-1)^2} + \frac{\alpha(1-\alpha)}{n}\right)}$$

Proposition 3.3

Lee (2013) formulated a symmetric truth detection model using unrelated questioning where some prior information about the mean of the study variable thus it uses the Bayesian method of estimation. This is because in the Bayesian method, the prior knowledge is used in the form of a prior distribution. When prior information is available in the form of a point guess, it can also be used in shrinking the estimator towards the prior point estimate.

To estimate the population mean of a sensitive quantitative variable (μ_A), we let $\hat{\mu}_{AO}$ be a prior estimator of μ_{AO} available from past study or simply an intelligent guess and k be the strength of the belief in the prior estimate of μ_{AO} . The estimator of μ_{AK} will be given by;

$$\hat{\mu}_{Ak} = k\mu_A + (1{-}k)\mu_{AO} \tag{3.11}$$

where $0 \leq k \leq 1$

If we let $d = \frac{\mu_{A0}-\mu_A}{\mu_A}$, then the bias of $\hat{\mu}_{Ak}$ will be given by;

$$\hat{\mu}_{Ak} = d(1{-}k)\mu_A \tag{3.12}$$

The mean squared error (MSE) of the estimator $\hat{\mu}_{Ak}$ is given by

$$MSE(\hat{\mu}_{Ak}) = E(\hat{\mu}_{Ak} - \mu_A)^2 = \frac{k^2\mu_A^2 U_A}{n} + d^2\mu_A^2(1-k)^2 \tag{3.13}$$

Lee (2013) formulated the model by letting T be the true response, R_1 and R_2 be the first and the second question respectively. The probability of answering the first question is p and the probability of answering the second question is $1 - p$.

A strategy is presented in which a respondent is instructed to answer "yes" if he or she belong to attribute "A" if not, he or she is required to draw a card from deck I of cards containing two statements:

1. I have A attribute with probability p

2. I do not have A attribute with probability $1 - p$

After drawing the card the respondent is required to answer "yes" or "no" accordingly without reporting the statement on the card to the interviewer. Also, the respondent proceed to next stage by answering "yes" if he or she belong to character "B". If not, he or she is required to draw another card from deck II of cards containing either of the two statements:

1. I have character B attribute with probability T

2. I do not have B attribute with probability 1–T

The respondent is then required to answer "yes" or "no"accordingly without reporting the statement on the card to the interviewer.

The expected responses will be four;

True Yes, False Yes, True No or False No.

Using the assumption of a known distribution of the variable B such that $\mu_B = 1$, δ_B^2= λ^2, $\lambda_i = 1$ if the i^{th} respondent is randomly assigned to the first statement in R_1, and $\lambda_i = 0$ if respondent is randomly assigned to the second statement in R_1. Further, α_i= 1 if the i^{th} respondent is randomly assigned to the first statement in R_2, and $\alpha_i = 0$ if the i^{th} respondent is randomly assigned to the second statement in R_2.

Proposition 3.3

Let $E(Y_i)$ be the expected value of the observed response and that $\hat{\mu}_{AO}$ be a prior estimate of $\mu_{AO,}$ then

$$\hat{\mu}_{AO} = \frac{E(Y_i) + (p-1)}{2p-1} \tag{3.14}$$

Proof

Let

$E(Y_i)$= $p\hat{\mu}_{AO}$ + (1-p)(1-$\hat{\mu}_{AO}$)

$E(Y_i)$= $p\hat{\mu}_{AO}$ + 1 - $\hat{\mu}_{AO} - p + p\hat{\mu}_{AO}$

$E(Y_i)$= 2$p\hat{\mu}_{AO}$ - $\hat{\mu}_{AO}$+ 1 − p

$E(Y_i)$= $\hat{\mu}_{AO}$(2p - 1) + 1-p

Making $\hat{\mu}_{AO}$ the subject of the formula the mean for$\hat{\mu}_{AO}$ is obtained as;

$$\hat{\mu}_{AO} = \frac{E(Y_i) - (p-1)}{2p-1} \tag{3.15}$$

This completes the proof.

Proposition 3.4

The variance of the mean estimator of $\hat{\mu}_T$ can is given as;.

$$Var(\mu_T) = \frac{1}{n}\mu_T^2 \left\{ \frac{\delta_T^2}{\mu_T^2} + (1 + \frac{\delta_T^2}{\mu_T^2}(1-p)(1-\alpha\lambda^2) \right\} \tag{3.16}$$

Proof

Let

$$Var(\hat{\mu}_T) = Var(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}Var(Y_i) \tag{3.17}$$

We know that;

$$Var(Yi) = E(Y_i^2) - E(Y_i)^2. \tag{3.18}$$

Now

$E(Y_i^2) = E(\alpha_i^2)E(T_i^2) + E(1+\alpha_i^2-2\alpha_i)\ \{E(\alpha_i^2)E(T_i^{\ 2}) + E(1-\alpha_i)^2E(T_i^2B_i^2)$

$+\ 2E(\alpha_i$- $\alpha_i{}^2)\mathrm{E}(T_i^2)\mathrm{E}(B_i)\} + 2\mathrm{E}\{\alpha_i(1-\alpha_i)\}\mathrm{E}\ (T_i^2)\mathrm{E}\{(\alpha_i + (1-\alpha_i)B_i\}$

$= p(\mu_T^2 + \delta_T^2) + (1-p)\{\alpha(\mu_{\mathrm{T}}^2 + \delta_T^2) + (1-\alpha)(\mu_T^2 + \delta_T^2)(\mu_B^2 + \delta_B^2) +$

$+2(\alpha-\alpha)(\mu_T^2 + \delta_T^2)$

$$= p(\mu_T^2 + \delta_T^2) + (1-p) \tag{3.19}$$

Therefore;

$Var(Yi) = p(\mu_T^2 + \delta_T^2) + (1-p)\{\alpha(\mu_T^2 + \delta_T^2) + (1-\alpha)(\mu_T^2 + \delta_T^2)(1 + \lambda^2)\} - \mu_T^2$

factoring out $\mu_T^2 + \delta_T^2$ we get

$= (\mu_T^2 + \delta_T^2)\{p + (1-p)\alpha + (1-p)(1-\alpha)(1+\lambda^2)\} - \mu_T^2$

$= (\mu_T^2 + \delta_T^2)\{1 + (1{-}p)(1{-}\alpha)\lambda^2\}$- μ_T^2

opening the brackets we get

$= (\mu_T^2{+}\delta_T^2){+}$ $(\mu_T^2{+}\delta_T^2)(1-p)(1-\alpha)\lambda^2{-}\mu_T^2$

so;

$$Var(Y_i) = (\delta_T^2 + (\mu_T^2 + \delta_T^2)(1-p)(1-\alpha)\lambda^2 \tag{3.20}$$

This implies that;

$$Var(\hat{\mu}_T) = \frac{1}{n}\left\{\delta_T^2 + (\mu_T^2 + \delta_T^2)(1-p)(1-\alpha\lambda^2\right\}$$

$$= \frac{1}{n}\mu_T^2\left\{\frac{\delta_T^2}{\mu_T^2} + (1 + \frac{\delta_T^2}{\mu_T^2}(1-p)(1-\alpha\lambda^2\right\} \tag{3.21}$$

This completes the proof.

If we let

$U_T{=}\left\{\frac{\delta_T^2}{\mu_T^2} + (1 + \frac{\delta_T^2}{\mu_T^2}(1-p)(1-\alpha\lambda^2\right\}$

The mean square error can be obtained as follows.

$$MSE(\hat{\mu}_T) = \sqrt{\frac{U_T}{n^2}} \tag{3.22}$$

3.1.2 Symmetric Truth detection Models

The symmetric truth detection model is an extension of the Asymmetric truth detection model. The model uses two randomization devices as opposed to Asymmetric truth detection model which uses one randomization device. The observable responses are no longer linked in any straightforward way to the respondent's true status and therefore the respondents are not safe in any way by denying the sensitive attribute. In this technique both "yes" and "no" responses are obtained from both guilty and innocent respondents. Thus, there is no possibility of playing safe by answering "no" and consequently no incentive to disregard the instruction. In such a symmetric design, interviewees not holding the sensitive attribute are expected to feel less uneasy when saying "yes" and be more likely to follow the RRT rules than in an asymmetric design.

The asymmetric response has two weaknesses namely;

i. large variance which may lead to misleading results.

ii. High level of non response.

In formulating symmetric truth detection models we have used two randomization devices, D_1 and D_2. The respondents are presented with the two randomization devices and are allowed to choose between the two devices. The respondents are then instructed to pick a card from the selected device and then respond truthfully to the question on the card. Each of the devices contains two types of cards; A bearing the statement "I belong to group A" and A^c bearing the statement "I do not belong to group A" selected using the simple random sampling with replacement. The results from these cards is used to estimate the proportion of individuals A who possess the sensitive attribute denoted as α. We also let a and b be any two positive integers; where a is the number of the respondents who choose the first randomization device and b is the number of the respondents who choose the second randomization device.

Further we let q be the probability of selecting the first randomization device D_1 and $1-q$ be the probability of selecting the second randomization device such that

$$q = \frac{a}{a+b}$$

and

$$1 - q = \frac{b}{a+b}$$

The effect of this model on the Asymmetric truth detection model is reducing the variance and non response thus improving the efficiency.

In the first randomization device $D_{1,}$the selection of the statement is done with probabilities p_1 and $1-p_1$ for selecting statement (i) and (ii) respectively. In the second randomization device D_2 the selection of the the statement is done with probabilities p_2 and $1 - p_2$ for statement (i) and (ii) respectively.

There are three conditions for this model;

i. $a + b \leq n$, where n is the sample size.

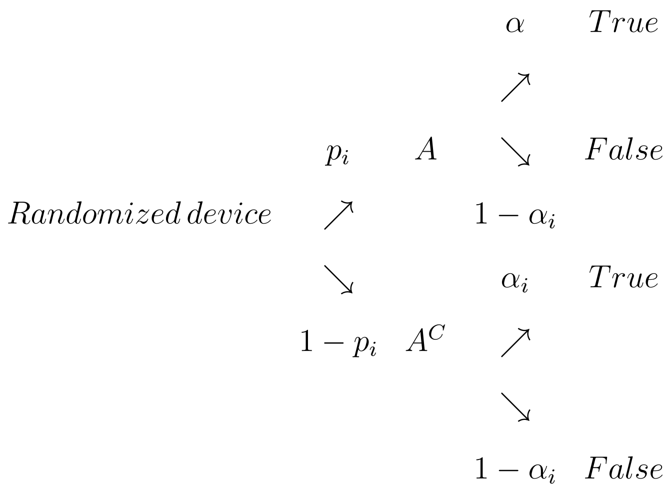
ii. $a \neq b$

iii. If $a - b$ is less than 0, then the absolute value is used.

To formulate the symmetric truth detection model, we have used the probability tree presented below.

Probability Tree Diagram

.



In the tree diagram above, p_i is the probability that a respondent is directed to answer the sensitive question and $1 - p_i$ is the probability that he or she is instructed to answer the non sensitive question, A is the group with sensitive attribute while A^c is the group without sensitive attribute. From elementary probability theory, the total proportion of “yes” answers irrespective of the question of affirmative response, can be expressed in the following way; p(True response) = p(the first question)P(presence of the sensitive attribute) + P(the second question)P(absence of the sensitive attribute).

The probability of "*yes*" response (λ) is therefore given as;

$$\lambda = \frac{a}{a+b}\left\{p_1\alpha + (1-p_1)(1-\alpha)\right\} + \frac{b}{a+b}\left\{p_2\alpha + (1-p_2)(1-\alpha)\right\} \quad (3.23)$$

Theorem 3.1

The unbiased estimator of the proportion of those with the sensitive attribute(α) is given by;

$$\hat{\alpha} = \frac{\hat{\lambda}(a+b) - p_1 b - p_2 a}{(2p_1 - 1)\,(a-b)}$$

Proof

From equation (3.23) we have;

$$\lambda = \frac{a}{a+b}\left\{p_1\alpha + (1-p_1)(1-\alpha)\right\} + \frac{b}{a+b}\left\{p_2\alpha + (1-p_2)(1-\alpha)\right\} \quad (3.24)$$

which can be simplified as;

$$\lambda = \frac{a\left\{p_1\alpha + (1-p_1)(1-\alpha)\right\} + b\left\{p_2\alpha + p_1(1-p_2)(1-\alpha)\right\}}{a+b} \quad (3.25)$$

which can further be simplified as;

$$\lambda = \frac{2p\alpha a + a - \alpha a - p_1 a + p_1 \alpha a + 2p_2 \alpha b + b - \alpha b - bp_2}{a+b} \tag{3.26}$$

After simplification, equation (3.26), reduces to;

$$\lambda = \frac{\alpha \left\{(2p_1 - 1)(a-b)\right\} + p_1 b + p_2 a)}{(a+b)} \tag{3.27}$$

Making α the subject of the formula, we get;

$$\alpha = \frac{\lambda(a+b) - p_1 b - p_2 a)}{\{(2p_1 - 1)(a-b)\}} \tag{3.28}$$

The unbiased estimator of α is therefore given by;

$$\hat{\alpha} = \frac{\hat{\lambda}(a+b) - p_1 b - p_2 a)}{\{(2p_1 - 1)(a-b)\}} \tag{3.29}$$

This complete the proof.

Theorem 3.2

The unbiased estimator of the variance of the existence of the sensitive characteristic (α) is given by;

$$V(\hat{\alpha}) = \frac{(p_2 a + p_1 b)(p_1 a + p_2 b)}{n(2p-1)^2(a-b)^2(a+b)^2} + \frac{\alpha(1-\alpha)}{n}$$

Proof

Using equation (3.29),

$$Var(\hat{\alpha}) = Var\left(\frac{\hat{\lambda}(a+b) - p_1 b - p_2 a)}{\{(2p_1-1)(a-b)\}}\right) \tag{3.30}$$

which can be simplified as;

$$Var(\hat{\alpha}) = Var\left(\frac{\hat{\lambda}(a+b)}{(2p_1-1)(a-b)}\right) - Var\left(\frac{p_1 b - p_2 a)}{(2p_1-1)(a-b)}\right) \tag{3.31}$$

But the variance of a constant is 0, therefore;

$$Var\left(\frac{p_1 b - p_2 a)}{(2p_1-1)(a-b)}\right) = 0$$

Hence;

$$Var(\hat{\alpha}) = \frac{Var(\hat{\lambda})(a+b)^2}{(2p_1-1)^2(a-b)^2}$$

$$= \frac{\frac{\lambda(1-\lambda)}{n}(a+b)^2}{(2p_1-1)^2(a-b)^2}$$

$$= \frac{\frac{\lambda(1-\lambda)}{n}(a+b)^2}{(2p_1-1)^2(a-b)^2} \tag{3.32}$$

From equation (3.27),

$$\lambda = \frac{\alpha\{(2p_1-1)(a-b)\} + p_1 b + p_2 a)}{(a+b)} \tag{3.33}$$

Substituting for λ in equation (3.32), we have;

$$Var(\hat{\alpha}) = \frac{[\alpha\left\{(2p_1-1)(a-b)\right\} + p_1 b + p_2 a)][(a+b) - \{\alpha(2p_1-1)(a-b)\} + p_1 b + p_2 a](a+b)^2}{n(2p_1-1)^2(a-b)^2(a+b)} \tag{3.34}$$

Which reduces to;

$$Var(\hat{\alpha}) = \frac{\alpha(1-\alpha)}{n} + \frac{2a\alpha(1-p_1-p_2)}{n(2p_1-1)^2(a-b)^2(a+b)} + \frac{(p_2 a - p_1 b)(p_1 a + p_2 b)}{n(2p_1-1)^2(a-b)^2(a+b)^2} \tag{3.35}$$

Since $p_1 = 1 - p_2$, then;

$$\frac{2a\alpha(1-p_1-p_2)}{n(2p_1-1)^2(a-b)^2(a+b)} = 0 \tag{3.36}$$

Thus;

$$Var(\hat{\alpha}) = \frac{\alpha(1-\alpha)}{n} + \frac{(p_2 a - p_1 b)(p_1 a + p_2 b)}{n(2p_1-1)^2(a-b)^2(a+b)^2} \tag{3.37}$$

This completes the proof.

3.2 Efficiency of Symmetric Truth Detection Model

In this chapter, we have performed the comparative study of the asymmetric and symmetric models mathematically and empirically. Mathematically, the basis for comparison is the variance.We wish to compare the variance of the symmetric truth detection models with the variance of asymmetric truth detection models.

3.2.1 Theoretical comparison

Mathematically, the basis for comparison of the asymmetric and Symmetric Truth Detection Models is the variance.We wish to test theoretically that the variance of the symmetric truth detection model is less than that of asymmetric Model. If the relative efficiency of the variance of the Symmetric Truth detection model (TDM) with respect to asymmetric TDM is greater than 1 (RE $>$ 1), then the Symmetric truth detection model will be more efficient . We wish to show that;

The theoretical comparison of our model was done with the earlier models which used a single randomization device. As both the asymmetric and symmetric models provided the unbiased estimator for variance ($\hat{\alpha}$), the Symmetric Truth Detection Models will be more efficient that the Asymmetric Truth Detection Model if;

Vary ($\hat{\alpha}_{Asy}$) - Vary ($\hat{\alpha}_{sy}$) >1.

We now wish to show the theoretical proof by finding variance difference between the two models.

$$Vary(\hat{\alpha}_{Asy}) - Vary(\hat{\alpha}_{sy}) = \frac{\alpha(1-\alpha)}{n} + \frac{p(1-p)}{n(2p-1)^2} - \frac{\alpha(1-\alpha)}{n} - \frac{(p_2a - p_1b)(p_2a + p_1b)}{n(2p_1-1)^2(a-b)^2(a+b)^2} \tag{3.38}$$

which can be simplified as;

$$Vary(\hat{\alpha}_{Asy}) - Vary(\hat{\alpha}_{sy}) = \frac{p(1-p)}{n(2p-1)^2} - \frac{(p_2a - p_1b)(p_2a + p_1b)}{n(2p_1-1)^2(a-b)^2(a+b)^2}$$

This can further be simplified as;

$$Vary(\hat{\alpha}_{Asy}) - Vary(\hat{\alpha}_{sy}) = \frac{p(1-p)}{n(2p-1)^2} - \frac{p_2^2a^2 + p_2ap_1b - p_1bp_2a - p_1^2b^2)}{n(2p_1-1)^2(a-b)^2(a+b)^2}$$

$$Vary(\hat{\alpha}_{Asy}) - Vary(\hat{\alpha}_{sy}) = \frac{p-p^2}{n(2p-1)^2} - \frac{p_2^2a^2 - p_1^2b^2}{n(2p_1-1)^2(a-b)^2(a+b)^2}$$

Since according to Lee (2013), $n(2p-1)^2 < n(2p_1-1)^2(a-b)^2(a+b)^2$ and $n(2p_1-1)^2(a-b)^2(a+b)^2$is extraordinarily large as it includes $(a+b)^2$. This implies that;

$$\frac{p-p^2}{n(2p-1)^2} > \frac{p_2^2a^2 - p_1^2b^2}{n(2p_1-1)^2(a-b)^2(a+b)^2}$$

and that;

$$\frac{p-p^2}{n(2p-1)^2}-\frac{p_2^2a^2-p_1^2b^2}{n(2p_1-1)^2(a-b)^2(a+b)^2}>1$$

Thus

$$RE=\frac{Variance\,of\;Asymmetrict\;TDM}{Variance\,of\;the\;symmetric\;TDM}>1$$

Hence Symmetric Truth Detection model is more efficient than Asymmetric Truth Detection model.

3.2.2 Empirical Comparison

In this section we have done empirical comparison of the variance of the Asymmetric TDM and Symmetric TDM through data simulation. This was done by setting $n = 10$, $p = 0.3$, $\quad p_1 = p_2 = 0.1$, $\alpha = 0.7$, $b = 2$, $|a - b|$ and $a \neq b$. Using these parameters, we have calculated the relative efficiency of Symmetric TDM with respect to Asymmetric TDM and presented the results in Table 1 below.

Table 3.1: Relative Efficiency

						Asymmetric TDM	Symmetric TDM	Relative
n	p	p_1	α	a	b	variance	Variance	Efficiency
10	0.3	0.1	0.7	3	2	0.1523	0.05381	2.83
10	0.3	0.1	0.7	4	2	0.1523	0.0297	5.12
10	0.3	0.1	0.7	5	2	0.1523	0.0245	6.22
10	0.3	0.1	0.7	6	2	0.1523	0.0229	6.65
10	0.3	0.1	0.7	7	2	0.1523	0.0222	6.86
10	0.3	0.1	0.7	8	2	0.1523	0.0218	6.99

The results in Table 3.1 shows that, the Symmetric Truth Detection Model is more efficient than the Asymmetric Truth Detection Model since all the RE > 0. It can be observed that efficiency increases with the increase as the difference between a and b increases.

4 CHAPTER FOUR: STRATIFIED TRUTH DETECTION MODELS

4.1 Stratified sampling

According to Daroga and Chaudhary (1989), in stratified sampling the population of N units is subdivided into L strata, the h^{th} strata having N_h units $(h = 1, 2, ..., L)$. These sub-populations are non overlapping so that they comprise the whole population such that;

N_1+ N_2 + N_3 + ... + N_h= N

A sample is drawn from each stratum independently, the sample size within the h^{th} being n_h $(h = 1, 2, ..., L)$ such that;

n_1+ n_2 + n_3 + ... + n_h= n.

In stratified sampling, the following notations are used;

N_h= Total number of units.

n_h = Number of units in sample

W_h = Stratum weight

α_{hj} = Value of the unit in thej^{th}in the h^{th}stratum.

α_h = Strata mean

According to Daroga and Chaudhary (1989), when a population of N units is divided into L strata the proportion of the population with the sensitive attribute per unit can be written as;

$$\alpha_{st} = \sum_{h=1}^{L}\sum_{j}^{N_h}\frac{\alpha_{hj}}{N} = \sum_{h=1}^{L}\frac{N_h\alpha_h}{N} = \sum_{h=1}^{L}W_h\alpha_h \tag{4.1}$$

where *st* stands for stratified.

An estimator for the variance of $\hat{\alpha}_{st}$ given as;
$V(\alpha_{st}) = \sum\limits_{h=1}^{L}\frac{N_h^2V(\alpha_h)}{N^2} = \sum\limits_{h}^{L}W_h^2V(\alpha_h)$

4.2 Asymmetric stratified truth detection models

The Asymmetric stratified truth detection models were developed by letting the population of size N be composed of L disjoint strata, the h^{th} stratum being of size N_h $(h = 1, 2, \ldots, L)$. From the h^{th} stratum a sample of size n_h $\left(n = \sum\limits_{h=1}^{L} n_h\right)$ are selected by Simple Random Sampling with replacement.

Lee (2013) used one randomization device to select his sample of respondents. The respondents are then subjected to five questions as stated below;

i. Do you have the sensitive attribute A?

ii. Do you have the non sensitive attribute $\bar{A}$?

iii. Do you have the unrelated attribute Q ?

iv. Do say “ yes”?

v. Do say “no”?

The probabilities of selecting each question are; p_{h1}, $p_{h2,}$ p_{h3}, p_{h4} and p_{h5} with $\sum\limits_{i=1}^{5}p_{hi} = 1$. The respondents responds to each question 1, 2 and 3 selected with probabilities; p_{h1}, $p_{h2,}and\ ph_3$ as “yes” or “no” according to his or her character. The respondents are also instructed to only say “yes” to question 4 and say “no” to question 5. Let λ_h be the probability of obtaining “yes” response, α be the sensitive population proportion and α_{yh} be the population proportion of unrelated character. Let n_h^* be the number of respondents who say yes among the n_h who were selected with SRSWR in stratum h, the proportion of yes response is;

$\hat{\lambda}_h = \frac{n_h^*}{n_h}$

Lee (2013) obtained the estimate of the population proportion with the sensitive character α as;

$$\hat{\alpha} = \sum_{h=1}^{L} W_h . \hat{\alpha_h} \tag{4.2}$$

where $W_h = \frac{N_h}{N}$

$$\hat{\alpha_h} = \frac{\hat{\lambda}_h - (p_{h2} + p_{h3}\alpha_{hy} + p_{h4})}{p_{h1} - p_{h2}} \qquad (p_{h1} \neq p_{h2}) \tag{4.3}$$

where $\hat{\alpha}_h$ is the estimator for the Strata mean.

substituting for $\hat{\alpha_h}$ in equation 1 we get;

$$\hat{\alpha} = \sum_{h=1}^{L} W_h \left[\frac{\hat{\lambda}_h - (p_{h2} + p_{h3}\alpha_{hy} + p_{h4})}{p_{h1} - p_{h2}} \right] \tag{4.4}$$

which is the estimator for the population proportion with the sensitive attribute.

He also proved that the variance of the estimator $(\hat{\alpha})$ for population with the sensitive attribute α is given by;

$$V(\hat{\alpha}) = \sum_{h}^{L} W_h^2 V(\hat{\alpha}_h)$$

But according to Lee (2013);

$$V(\hat{\alpha}_h) = \frac{\alpha_h(1-\alpha_h)}{n_h} + \frac{\alpha_h(1-p_{h1}-p_{h2}-2p_{h3}\alpha_{hy}-2p_{h4})}{n_h(p_{h1}-p_{h2})}$$

$$+\frac{(p_{h2}+p_{h3}\alpha_{hy}+p_{h4})(1-p_{h2}-p_{h3}\alpha_{hy}-2p_{h4})}{n_h(p_{h1}-p_{h2})^2} \tag{4.5}$$

substituting for $V\ (\hat{\alpha_h})$ we obtain;

$$V(\hat{\alpha}) = \sum_{h=1}^{L} W_h^2 [\frac{\alpha_h(1-\alpha_h)}{n_h} + \frac{\alpha_h(1-p_{h1}-p_{h2}-2p_{h3}\alpha_{hy}-2p_{h4})}{n_h(p_{h1}-p_{h2})}$$

$$+\frac{(p_{h2}+p_{h3}\alpha_{hy}+p_{h4})(1-p_{h2}-p_{h3}\alpha_{hy}-2p_{h4})}{n_h(p_{h1}-p_{h2})^2}] \tag{4.6}$$

which can be simplified as;

$$V(\hat{\alpha}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} [\alpha_h(1-\alpha_h) + \frac{\alpha_h(1-p_{h1}-p_{h2}-2p_{h3}\alpha_{hy}-2p_{h4})}{(p_{h1}-p_{h2})}$$

$$+\frac{(p_{h2}+p_{h3}\alpha_{hy}+p_{h4})(1-p_{h2}-p_{h3}\alpha_{hy}-2p_{h4})}{(p_{h1}-p_{h2})^2}] \tag{4.7}$$

4.3 Symmetric stratified truth detection model

The Asymmetric stratified truth detection model has however several weaknesses. First, the questioning technique seems to force the respondent to admit having an attribute which he/she actually does not have. This is likely to lead to a high level of non response since the respondents may not be willing to admit an attribute they do not possess. Secondly, the model leads to a very large variance which is likely to give misleading results.

To overcome these weaknesses, we have developed a model which uses two samples as opposed to Asymmetric stratified TDM which use only one sample. The model also uses only two questions and does not include the forced response question. We have called this model the Symmetric Stratified Truth Detection Model.

In formulating the model the two samples are selected using two randomization devices namely; D_1 and D_2 respectively. The respondents are presented with the two randomization devices meant to protect their privacy and then they are allowed to choose one of the devices. The respondents who select the first randomization device are taken as sample one while those who select the second randomization device are taken as sample two. This is done by letting a and b be any two positive numbers; where a is the number of the respondents who choose the first randomization device and b is the number of the respondents who choose the second randomization device. There are three assumptions for this study, namely;

i. $a + b \leq n$.

ii. The $a + b$ is old so that, $a \neq b$.

iii. For $a - b < 0$, absolute value is used.

Further we let q be the probability of selecting the first randomization device D_1 and $1 - q$ be the probability of selecting the second randomization device such that;

$$q = \frac{a}{a + b}$$

and

$$1 - q = \frac{b}{a + b}$$

After selecting the device, the respondents are required to respond to two questions as opposed to the Asymmetric stratified truth detection model where the respondents were presented with five questions with two of them being forced responses. These questions presented as follows;;

i. Do you belong to group A (say Yes or No)

i. Do you belong to group $\bar{A}$ (Say Yes or No)

Where group A represents those respondents who have the sensitive attribute while $\bar{A}$ are those who are not having sensitive attribute. Let the population of size N be composed by L disjoint strata, the h^{th} stratum being of size N_h $(h = 1, 2, \ldots, L)$. From the h^{th} stratum, respondents of $n_h \left(n = \sum_{h=1}^{L} n_h \right)$ are selected by SRSWR. In each stratum of size n_h respond are selected using the a random device D_{hi} $(h = 1, 2, \ldots, L,\ i = 1, 2)$ of the two sample. The respondents responds to each question 1 and 2 selected with probabilities; $p_{hi1} and\, p_{hi2}$ as "yes" or "no" selected according to his or her character. In the first randomization device $D_{1,}$the selection of the statement (i) or (ii) is done with probabilities p_{hi1} and $1 - p_{hi1}$ respectively. In the second randomization device D_2 the selection of the statement is done with probabilities p_{hi2} and $1 - p_{hi2}$ for statement (i) and (ii) respectively. Let α_h be the individuals who possess the sensitive attribute and λ_{hi} be the probability of "*yes*" response.

From equation 8 in chapter three, the probability of "*yes*" response (λ) is given by;

$$\lambda = \frac{a}{a+b} \left\{ p_1 \alpha + (1 - p_1)(1 - \alpha) \right\} + \frac{b}{a+b} \left\{ p_2 \alpha + (1 - p_2)(1 - \alpha) \right\} \quad (4.8)$$

for symmetric truth detection model. However when stratified sampling is used, the probability of "*yes*" response is given by;

$$\lambda_h = \frac{a}{a+b}\left\{p_{hi1}\alpha_h + (1-p_{hi1})(1-\alpha_h)\right\} + \frac{b}{a+b}\left\{p_{hi2}\alpha_h + (1-p_{hi2})(1-\alpha_{hi2})\right\} \tag{4.9}$$

Theorem 4.1

The unbiased estimator of the proportion of those with sensitive characteristic (α) is given by;

$$\hat{\alpha} = \sum_{h=1}^{L} W_h \left[\frac{\hat{\lambda}_h(a+b) - p_{hi1}b - p_{hi2}a}{(2p_{hi1}-1)\,(a-b)}\right] \tag{4.10}$$

where W_h is the weight for stratum h, p_{hi1} , $1-p_{hi1}$, p_{hi2}, and $1-p_{hi2}$ are as earlier explained.

Proof

From equation (4.1) we observed that;

$\alpha_{st}= \sum\limits_{h=1}^{L}\sum\limits_{j}^{N_h}\frac{\alpha_{hj}}{N} = \sum\limits_{h=1}^{L}\frac{N_h\alpha_h}{N} = \sum\limits_{h=1}^{L} W_h\ \alpha_h\ \alpha_h$, where j denotes the sampling unit while h denotes the stratum.

This implies that;

$$\hat{\alpha}_{st} = \sum_{h}^{L} W_h \hat{\alpha_h} \tag{4.11}$$

We therefore need to compute $\hat{\alpha_h}$

From equation (4.9),

$$\lambda_h = \frac{a}{a+b}\left\{p_{hi1}\alpha_h + (1-p_{hi1})(1-\alpha_h)\right\} + \frac{b}{a+b}\left\{p_{hi2}\alpha_h + (1-p_{hi2})(1-\alpha_h)\right\}$$

which can be expanded as;

$$\lambda_h = \frac{a\left\{p_{hi1}\alpha_h + (1-p_{hi1})(1-\alpha_h)\right\} + b\left\{p_{hi2}\alpha_{h2} + p_{hi1}(1-p_{hi2})(1-\alpha_h)\right\}}{a+b}) \tag{4.12}$$

After simplification, equation (4.12) reduces to;

$$\lambda_h = \frac{\alpha_h\left\{(2p_{hi1}-1)(a-b)\right\} + p_{hi1}b + p_{hi2}a)}{(a+b)} \tag{4.13}$$

Making α_h the subject of the formula, we get;

$$\alpha_h = \frac{\lambda_h(a+b) - p_{hi1}b - p_{hi2}a)}{\left\{(2p_{hi1}-1)(a-b)\right\}} \tag{4.14}$$

the unbiased estimator of α_h is therefore given by;

$$\hat{\alpha}_h = \frac{\hat{\lambda}_h(a+b) - p_{hi1}b - p_{hi2}a)}{\left\{(2p_{hi1}-1)(a-b)\right\}} \tag{4.15}$$

where $\hat{\lambda}_h$ is the unbiased estimator for λ_h.

But

$$\hat{\alpha}_{st} = \sum_{h=1}^{L} W_h \hat{\alpha}_h \tag{4.16}$$

Substituting for $\hat{\alpha}_h$ in equation (4.16) we get;

$$\hat{\alpha}_{st} = \sum_{h=1}^{L} W_h \left[\frac{\hat{\lambda}_h(a+b) - p_{hi1}b - p_{h2}a)}{\{(2p_{hi1}-1)(a-b)\}} \right]$$

This complete the proof.

.

.

Theorem 4.2

The variance for the estimator of those with sensitive characteristic (α_{st}) is given by;

$$Var(\hat{\alpha}_{st}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left[\alpha_h(1-\alpha_h) + \frac{(p_{hi2}a - p_{hi1}b)(p_{hi1}a + p_{hi2}b)}{(2p_{hi1}-1)^2(a-b)^2(a+b)^2} \right]$$

where the symbols are as earlier explained.

Proof

Using equation (4.15) above, the variance of α_h is estimated as;

$$Var(\hat{\alpha}_h) = Var\left(\frac{\lambda_{hi}(a+b) - p_{hi1}b - p_{hi2}a)}{\{(2p_{hi1}-1)(a-b)\}} \right) \tag{4.17}$$

Which simplifies to;

$$Var(\hat{\alpha}_h) = Var\left(\frac{\lambda_{ih}(a+b)}{(2p_{hi1}-1)(a-b)} \right) - Var\left(\frac{p_{hi1}b - p_{hi2}a)}{(2p_{hi}-1)(a-b)} \right) \tag{4.18}$$

But the variance of a constant is 0, therefore;

$$Var\left(\frac{p_{hi1}b - p_{hi2}a)}{(2p_{hi1}-1)(a-b)}\right) = 0$$

Equation (4.18), therefore reduces to;

$$Var(\alpha_h) = \frac{Var(\lambda_{hi})(a+b)^2}{(2p_{hi1}-1)^2(a-b)^2}$$

Substituting for Var(λ_{hi}) and λ_{hi} we have;

$$Var(\hat{\alpha_h}) = \frac{[\alpha_h\left\{(2p_{hi1}-1)(a-b)\right\} + p_{hi1}b + p_{hi2}a)][(a+b)}{n_h(2p_{hi1}-1)^2(a-b)^2(a+b)}$$

$$-\frac{\{\alpha_h(2p_{hi1}-1)(a-b)\} + p_{hi1}b + p_{hi2}a](a+b)^2}{n_h(2p_{hi1}-1)^2(a-b)^2(a+b)} \tag{4.19}$$

Which reduces to;

$$Var(\alpha_h) = \frac{\alpha_h(1-\alpha_h)}{n_h} + \frac{2a\alpha_h(1-p_{hi1}+p_{hi2})}{n_h(2p_{hi1}-1)^2(a-b)^2(a+b)} + \frac{(p_{hi1}a - p_{hi2}b)(p_{hi1}a + p_{hi2}b)}{n_h(2p_{hi1}-1)^2(a-b)^2(a+b)^2} \tag{4.20}$$

but;

$$\frac{2a\alpha_h(1-p_{hi1}+p_{hi2})}{n_h(2p_{hi}-1)^2(a-b)^2(a+b)} = 0 \tag{4.21}$$

Equation (4.21) therefore simplifies;

$$Var(\hat{\alpha}_h) = \frac{\alpha_h(1-\alpha_h)}{n_h} + \frac{(p_{i1}a - p_{hi2}b)(p_{hi1}a + p_{hi2}b)}{n_h(2p_{hi1}-1)^2(a-b)^2(a+b)^2} \tag{4.22}$$

Thus;

$$V(\hat{\alpha}_{st}) = \sum_{h}^{L} W_h^2 \left[\frac{\alpha_h(1-\alpha_h)}{n_h} + \frac{(p_{i1}a - p_{hi2}b)(p_{hi1}a + p_{hi2}b)}{n_h(2p_{hi1}-1)^2(a-b)^2(a+b)^2}\right]$$

Which simplifies as;

$$Var(\hat{\alpha}_{st}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left[\alpha_h(1-\alpha_h) + \frac{(p_{hi2}a - p_1b)(p_{hi1}a + p_{hi2}b)}{n_h(2p_{hi1}-1)^2(a-b)^2(a+b)^2}\right]$$

This completes the proof.

4.4 Theoretical comparison

The theoretical comparison of symmetric stratified truth detection model was done with the earlier models namely Asymmetric stratified truth detection model which used a single randomization device. As both the asymmetric and symmetric models provided the unbiased estimator for variance $(\hat{\alpha})$, the proposed model will be more efficient that the single randomized device model if the the variance of Asymmetric stratified truth detection model is greater than the variance of symmetric stratified truth detection. The proof was presented below;

$$V(\hat{\alpha}) = [\sum_{h=1}^{L}\frac{W_h^2}{n_h}[\alpha_h(1-\alpha_h) + \frac{\alpha_h(1-p_{h1}-p_{h2}-2p_{h3}\alpha_{hy}-2p_{h4})}{(p_{h1}-p_{h2})}$$

$$+\left[\frac{(p_{h2}+p_{h3}\alpha_{hy}+p_{h4})(1-p_{h2}-p_{h3}\alpha_{hy}-2p_{h4})}{(p_{h1}-p_{h2})^2}\right]$$

$$-\left[\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[\alpha_h(1-\alpha_h) + \frac{(p_{hi2}a-p_1b)(p_{hi1}a+p_{hi2}b)}{n_h(2p_{hi1}-1)^2(a-b)^2(a+b)^2}\right]\right]$$

This simplifies as;

$$\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[\frac{\alpha_h(1-p_{h1}-p_{h2}-2p_{h3}\alpha_{hy}-2p_{h4})}{(p_{h1}-p_{h2})} + \frac{(p_{h2}+p_{h3}\alpha_{hy}+p_{h4})(1-p_{h2}-p_{h3}\alpha_{hy}-2p_{h4})}{(p_{h1}-p_{h2})^2}\right]$$

$$-\left[\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[\frac{(p_{hi2}a-p_{hi1}b)(p_{hi1}a+p_{hi2}b)}{n_h(2p_{hi1}-1)^2(a-b)^2(a+b)^2}\right]\right]$$

which further simplifies as;

$$\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[\frac{\alpha_h(p_{h1}-p_{h2})(1-p_{h1}-p_{h2}-2p_{h4}) + (p_{h2}+p_{h3}\alpha_{hy}+p_{h4})(1-p_{h2}-p_{h3}\alpha_{hy}-2p_{h4})}{(p_{h1}-p_{h2})^2}\right]$$

$$-\left[\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[\frac{(p_{hi2}^2a^2-p_{hi1}^2b^2)}{n_h(2p_{hi1}-1)^2(a-b)^2(a+b)^2}\right]\right]$$

Since;

$$(p_{h1} - p_{h2})^2 < n_h \left(2p_{hi1} - 1\right)^2 (a-b)^2 (a+b)^2$$

it implies that;

$$\sum_{h=1}^{L} \frac{W_h^2}{n_h} \left[\frac{\alpha_h(p_{h1} - p_{h1})(1 - p_{h1} - p_{h2} - 2p_{h4}) + (p_{h2} + p_{h3}\alpha_{hy} + p_{h4})(1 - p_{h2} - p_{h3}\alpha_{hy} - 2p_{h4})}{(p_{h1} - p_{h2})^2} \right]$$

$$> \left[\sum_{h=1}^{L} \frac{W_h^2}{n_h} \left[\frac{(p_{hi2}^2 a^2 - p_{hi1}^2 b^2)}{n_h(2p_{hi1} - 1)^2(a-b)^2(a+b)^2} \right] \right]$$

then;

$$Var(\hat{\alpha}_{Ast}) > Var(\hat{\alpha}_{sst})$$

Hence Symmetric stratified Truth Detection Model is more efficient than the Asymmetric stratified Truth Detection Model.

4.5 Empirical Comparison

The relative efficiency (RE) of the Symmetric stratified truth detection model technique with respect to Asymmetric stratified was obtained by dividing the variance of the Asymmetric stratified truth detection model by the variance of Symmetric stratified technique after obtaining data using simulation method. The Symmetric stratified truth detection technique will be more efficient than

Symmetric stratified truth detection model if;

$$RE = \frac{Variance\ for\ Asymmetric\ stratified\ technique}{Variance\ for\ Symmetric\ stratified\ technique} > 0$$

We have done the comparison of the variance of the Asymmetric stratified technique and Symmetric stratified technique. For convenience of comparison, we assume the number of strata is two and that sample size is $n = 10$, with n_1= 7 and n_2=3, W_1=0.7 and W_2= 0.3. Without loss of generality we assume the selection probabilities are the are as follows;

p_{h11}= p_{h21}= 0.1, p_{h12}= p_{h22}= 0.2, = p_{h13}= p_{h23}= 0.3 and p_{h14}= p_{h24}= 0.2, $\alpha_h = \alpha_1 = \alpha_2 = \alpha_{1y} = \alpha_{2y} = 0.7$, $b = 2$, 3≤a≤ 8, $a - b > 0$ and $a \neq b$. Using these parameters, we have calculated the relative efficiency of Symmetric stratified TDM with respect to Asymmetric stratified TDM and presented the results in Table 4.1 below.

Table 4.1: Relative Efficiency

		Asymmetric stratified TDM	Symmetric stratified TDM	Relative
a	b	variance	Variance	Efficiency
3	2	0.07888	0.031552	2.5
4	2	0.07888	0.01753	4.5
5	2	0.07888	0.01461	5.4
6	2	0.07888	0.01337	5.9
7	2	0.07888	0.01289	6.12
8	2	0.07888	0.01264	6.24

The results in Table 4.1 shows that, the Symmetric stratified truth detection model yields relatively less variance compared to Asymmetric stratified truth detection model and therefore more efficient since the RE > 1. It can be observed that efficiency increased with the increase in the difference between a and b.

4.6 Real life application

4.6.1 Data collection

Data for this study was collected from form three students in Muthale Girls' Secondary Schools in Kitui County. The symmetric variant truth detection model and symmetric stratified truth detection models were applied in testing the examination cheating among the secondary school students so as to validate the models.

To validate the symmetric models, we have used one randomization device as well as two randomization devices and compared the results. For the two randomization devices we have used, park A and Park B as device D_1 and D_2 respectively. The participants were 260 students composed of Form three students. From these students, 130 were subjected to Asymmetric randomization approach while 130 were subjected to symmetric randomization. In the asymmetric questioning baseline condition, respondents were presented with three questions while in the symmetric approach they were presented with two questions and instructed to answer accordingly. For the symmetric questioning, the respondents were first to select a device, pick a card from the device and answer accordingly.

The survey data obtained by symmetric approach were compared with the corresponding information obtained by the symmetric approach to test the research hypotheses below;

H_{A1} :When facing ethical decisions concerning students examinations, the proportion of the students reporting decisions in line with the statement of responsibility is less for those receiving the symmetric questionnaire than those receiving Asymmetric questionnaire.

H_{A2} :The overall response rate is less for the students receiving the symmetric approach questioning than for the students receiving the Asymmetric approach questionnaire.

The Asymmetric questioning was as follows;

1. What is the type of your admission number?

(a) Even number ( )

(b) Odd number ( )

2. If your admission number is even number, then please answer truthfully to the question below but if it is odd then ignore the question and simply say “yes”.

i. Have you ever cheated in examination? Yes ( ) No ( )

3. If your admission number is odd number, then please answer truthfully to the question below but if it is even then ignore the question and simply say “yes”.

i. Have you ever cheated in examination? Yes ( ) No ( )

The questions for the symmetric approach were as follows;

1. What randomization device did you choose?

(a) D_1 ( )

(b) D_2 ( )

2. Have you ever cheated in any examination? Yes ( ) No ( )

For stratification, there were two stratum, those who selected D_1 and those who selected D_2, making two groups.

The data from Asymmetric and Symmetric questioning was computed and compared to check the validity of the symmetric approach method in testing examination dishonesty. Since the actual instances of the unethical decisions is not known, it was assumed that higher instances of unethical decisions reflect more honest responses. Thus symmetric results that indicate students cheat in exams significantly more times than do Asymmetric results are evidence that symmetric approach can be used to obtain more reliable data in this study.

Data for this study was also collected using stratified methods for both the asymmetric questioning and symmetric questioning. The symmetric variant truth detection model was applied in testing the examination dishonesty among form three and four he secondary school students. The respondents were form three and form four students.

4.6.2 Data analysis

Results from Table 4.2 shows that the number of “yes”responses for secondary students using Symmetric questioning was significantly higher (76.1%) compared to the “yes” responses under Asymmetric questioning which were 56.1%. The same results were revealed by Table 5.3, where the “yes” responses for symmetric stratified questioning were significantly higher (60.7%) compare to the Asymmetric questioning which had 41.5%.

Table 4.2: Unstratified student's Results

	$Asymmetric\,Questioning$	$Symmetric\,Questionning$
ni	130	130
$n(Yes)$	73(56.1%)	99(76.1%)
$n(No)$	57(43.9%)	31(23.9%)
		$n = 260$

Table 4.3: Stratified student's Results

	$Asymmetric\,stratified\,Questioning$	$Symmetric\ stratified\ Questionin$
N	130	130
$n(Yes)$	54(41.5%)	79(60.7%)
$n(No)$	76(58.5%)	51(39.3%)
		$n = 260$

From the results in Table 4.3; the following results were obtained for Asymmetric sampling; $\lambda = 0.561$ which is the observed proportion of "yes", $p = 0.331$ is the probability that a respondent is directed to answer the sensitive question and 0.569 is the probability that the respondent is instructed to answer the non sensitive question. Let α be the probability of the existence of the sensitive attribute and 1- α be the probability of non existence of the sensitive attribute.

From equation 2.1; the proportion of sensitive attribute α is estimated as;

$$\hat{\alpha} = \frac{\hat{\lambda} + (p - 1)}{2p - 1}$$

substituting for $\lambda$and p we get;

$$\hat{\alpha} = \frac{0.561 + (0.331 - 1)}{2(0.331) - 1} = 0.32$$

From equation (2.2), the variance of the estimator of α is given by;

$$V(\hat{\alpha}) == \frac{\alpha(1 - \alpha)(}{n} + \frac{p(1 - p)}{n(2p - 1)^2}$$

Substituting for the unknown values, we have;

$$V(\hat{\alpha}) = \frac{0.32(1 - 0.32)}{260} + \frac{(0.331)(1 - 0.331)}{260(2(0.331) - 1)^2} = 0.00822$$

The variance for the symmetric truth detection model was given as;

$$Var(\hat{\alpha}_{Asy}) = \frac{\alpha(1 - \alpha)}{n} + \frac{(p_2a - p_1b)(p_1a + p_2b)}{n(2p_1 - 1)^2(a - b)^2(a + b)^2}$$

To calculate the variance of the symmetric stratified truth detection model, the following parameters were used;

$n = 260,\ a = 182,\ b = 78, \alpha = 0.32, p_1 = 0.7, p_2 = 0.3,$

$$Var(\hat{\alpha}) = \frac{0.32(1 - 0.32)}{260} + \frac{(0.3(182) - 0.3(78))(0.7(182) + 0.3(78))}{260(2(0.7) - 1)^2(182 - 78)^2(182 + 78)^2}$$

$$Var(\hat{\alpha}_{Sy}) = 0.0008578$$

Hence the variance for symmetric truth detection model $<$ symmetric truth detection model

and this makes symmetric truth detection model it more efficient than Asymmetric truth detection model.

Application of stratified;

We have calculated the Asymmetric stratified truth detection variance using the models below as previously proved.

$$W_h = \frac{N_h}{N},\ \hat{\alpha} = \sum_{h=1}^{L} W_h.\hat{\alpha_h},\ \hat{\alpha}_h = \frac{\hat{\lambda}_h - (p_{h2} + p_{h3}\alpha_{hy} + p_{h4})}{p_{h1} - p_{h2}}$$

$$\hat{\alpha}_h = \frac{0.561 - (0.2 + 0.2(0.7) + 0.2)}{0.3 - 0.2} = 0.21$$

From Table 3, we have the following results for Asymmetric stratified approach;

Also;$n_1 = n_2 = 130$,$\lambda_1 = \lambda_2 = 0.42$, $W_1 = W_2 = 0.7$ p_{h11}= p_{h21}= 0.3, p_{h12}= p_{h22}= 0.2, = p_{h13}= p_{h23}= 0.2 and p_{h14}= p_{h24}= 0.2, $\alpha_h = \alpha_1 = \alpha_2 = \alpha_{1y} = \alpha_{2y} = 0.21$, $a = 137$, $b = 123$

To obtain the variance of Asymmetric stratified truth detection model we used equation 5.7 which was given as;

$$V(\hat{\alpha}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}[\alpha_h(1-\alpha_h) + \frac{\alpha_h(1-p_{h1}-p_{h2}-2p_{h3}\alpha_{hy}-2p_{h4})}{(p_{h1}-p_{h2})}$$

$$+\frac{(p_{h2}+p_{h3}\alpha_{hy}+p_{h4})(1-p_{h2}-p_{h3}\alpha_{hy}-2p_{h4})}{(p_{h1}-p_{h2})^2}]$$

Substituting the above parameters in equation we obtained;

$$V(\hat{\alpha}) = 2[\frac{0.49}{130}[0.21(1-0.21) + \frac{0.21(1-0.3-0.2-2(0.2)(0.21)-2(0.2))}{(0.3-0.2)}$$

$$+\frac{(0.2+0.2(0.21)+0.2)(1-0.2-0.2(0.21)-2(0.2))}{(0.3-0.2)^2}] = 0.2812$$

The symmetric stratified truth detection variance was calculated by substituting for unknowns in equation 23 given as;

Also;$n_1 = n_2 = 130$,$\lambda_1 = \lambda_2 = 0.42$, $W_1 = W_2 = 0.7$ p_{h11}= p_{h21}= 0.3, p_{h12}= p_{h22}= 0.2, = p_{h13}= p_{h23}= 0.2 and p_{h14}= p_{h24}= 0.2, $\alpha_h = \alpha_1 = \alpha_2 = \alpha_{1y} = \alpha_{2y} = 0.21$, $a = 182$, $b = 78$.

Equation 23 was presented as;

$$Var(\hat{\alpha}_{st}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left[\alpha_h(1-\alpha_h) + \frac{(p_{hi2}a - p_{hi1}b)(p_{hi1}a + p_{hi2}b)}{n_h(2p_{hi1}-1)^2(a-b)^2(a+b)^2} \right]$$

$$Var(\hat{\alpha}_{st}) = 2(\frac{0.49}{130} \left[0.21(1-0.21) + \frac{(0.2(182) - 0.3(78))(0.3(182) + 0.2(78))}{130(2(0.3)-1)^2(182-78)^2(182+78)^2} \right] = 0.1228$$

It can be observed from the two results that the variance for symmetric stratified truth detection model is relatively less than the symmetric stratified truth detection model.

Since the actual instances of the examination dishonesty is not known, it was assumed that higher instances of yes responses reflect more honest responses. Thus RR results that indicate students make unethical decisions were significantly more than the direct query results. This was taken as an evidence that RR can be used to obtain more reliable data in this study. Both studies established that the “yes” responses under the RRT were more in both cases. This is an indication that RRT is a valid and reliable test compared to direct questioning. It was also established that the symmetric variate truth detection model offered more reliable results compared to asymmetric model. We do therefore reject $H_A$which stated that when facing ethical decisions concerning students examinations dishonesty, the proportion of the students reporting decisions consisted with the statement of responsibility in examination taking is less in the RR questionnaire than those receiving direct questionnaire and conclude that the overall response rate is less for the students receiving the direct questioning than for the students receiving the RR questionnaire.

5 CHAPTER FIVE: CONCLUSION, RECOMMENDATIONS AND FUTURE RESEARCH

In this thesis we have formulated Symmetric and Stratified truth detection models using randomized response technique which can be used in investigating sensitive information such as; rape, abortion, tax evasion and examination dishonesty, among others. This work improves the relative efficiency of the classical randomized response technique of the Asymmetric truth detection models and the Asymmetric stratified truth detection models. Many researchers such as the 2019 Violence Against Children Survey (VACS) and the 2018/2019 Kenya Population based HIV impact assessment (KENPHIA), collected self report information on social demographic characteristics of respondents and self report disease status of some of them. The extended Symmetric RRT formulated in this study can be used in complex surveys to determine the status some of the self reported indicators among respondents. It is of our great interest if part of future work could include Simulation and Extrapolation Exercise (SIMEX) misclassification models within the survey context. This work reverberates adjustment for misclassification in survey methods. We do therefore recommend researchers on sensitive information to use symmetric truth detection models as they yield more reliable data compared to the Asymmetric model truth detection models .

5.1 Further Research

In this study we have formulated symmetric truth detection model and symmetric stratified truth detection model using a randomized approach. Further research can be done on the following:

1. How the randomized response can be used in controlling sampling errors.

2. The optimum estimation of responses in the randomized response situations.

3. Estimation of the correlation of responses among various groups using randomized response technique.

6 REFERENCES

Adebola F.B & Johnson N.A (2015). A modified Stratified randomized response. Mathematical theory and modeling 4(13), 29-42.

Adepetun, A.O. and Adebola, F.B. (2014): On the Relative Efficiency of the Proposed Reparametized Randomized Response Model. International Journal of Mathematical Theory and Modeling, 4, 58-67.

Ahn, S. C., Lee, G. S.(2003). A Stratified Unrelated Question Model. Journal of the Korean Data Analysis Society, 5(4B), (2003), 853-864.

Antonak, R. F. & Livneh, H. (1995). Randomised response technique: A review and proposed extension to disability attitude research. Genetic, Social, and General Psychology Monographs121: 97–145.

Bar-Lev SK, Bobovich E, Boukai B (2003). A Two-Stage Sequential Scheme for Warner's Response Model. Communications in Statistics - Theory and Methods 32(12) 2375-2389.

Bhargava, M., Singh, R.A (2000). Modified Randomization Device for Warner's Model. Statistica, 60, (2000),315-321. bias.

Boeije, H. & Lensvelt-Mulders, G. J. L. M. (2002). Honest by chance: A qualitative interview. Boruch Journal of the American Statistical Association 60: 63–69.

Bo Yu, Zongda J., Jiayong T. and Ge G. (2015): Estimation of sensitive information by randomized response data in successive sampling, 4(13), 29-42.

Chang, H. J., Wang, C. L., Huang, K. C. (2004) On Estimating the Proportion of a Qualitative Sensitive Character Using Randomized Response Sampling, Quality and Quantity. Journal of the American Statistical Association 60: 63–69. 38,675-680.

Chaudhuri, A. and Mukerjee, R. (1988). Randomized response: Theory and Techniques, Marcel Dekker, Inc.,NewYork..

Clark, S. J. & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting datacollection, perception of risks and losses, and motivation to give truthful answers to Design. Communications in Statistics-Theory and Methods, 5, (1976),565-574.

Christopher, R. (2019). A new Randomized Response Model. Journal of the Royal Statistical Society,68(3): 523-530

Eichhorn B.H., Hayre L.S. (2003). Scrambled Randomized Response Methods for Obtaining Sensitive Quantitative Data. J of Statistical planning and Inference 7:307–316.

Fox, J. A. & Tracy, P. E.(1986).Randomised Response: A Method for Sensitive Surveys. Beverly Hills: Sage Publications

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

Guerriero, M., & Sandri, M. F. (2007). A note on the comparison of some randomized response procedures. Journal of statistical planing and inference, 173, 2184-2190.

Heijden, P. G. M. v. d., Gils, G. v., Bouts, J. & Hox, J. (1998). A comparison of randomised Horvitz, D. G., Greenberg, B. G., Abernathy, J. Randomized Response: a Data-Gathering Device for Sensitive Question, International Statistical Review, 44,.(1976), 181-196.

Horvitz, D. G., Shah, B. V. & Simmons, W. R. (1967). The unrelated question randomised response model. Proceedings of the Social Statistics Section, ASA: 65–72.

Jarman,B. J. (1997). The Prevalence and Precedence of Socially Condoned Sexual agression Journal of the American Statistical Association 66(335): 627–629.

Kendall, M. K. & Stuart, A. (1979). The Advanced Theory of Statistics, Vol. 2. New York:

Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. Biometrika 77(2): 436–438.

Lanke, J. (1976). On the degree of protection in randomized interviews. International Statistical Review, 44, 197–203.

Lee, G. S. (2012). Unrelated Question Randomized Response Model by Stratified Replicated Systematic Sampling, Journal of the Korean Data Analysis Society, 14(2B),(2010), 781-791.

Lee, G.S (2015). Estimating at least seven measures of qualitative variable from a single sample using randomized response. Statistics and prob letters 83, 399-409.

Lee, G. S., Hong, K. H. (1999). Modified Nayak's Randomized Response Model, Communications of Korean Statistical Society, 6(1), 117-129. (in Korean).

Lensvelt-Mulders, G.J.L.M.; Hox, J.J.; van der Heijden, P.G.M.; Maas, C.J.M. (2005). Meta- Leysieffer, F. W., Warner, S. L. Respondent Jeopardy and Optimal Designs in RR Models,Journal of the American Statistical Association, 72, (1976) 649-656.

Ljungqvist, L. (1993). A unified approach to measures of privacy in randomized response models: A utilitarian perspective. Journal of the American Statistical Association, 88, 97–103.

Lyons, W. (1999). "Early Voting and the Timing of the Vote: Unanticipated Consequences of Electoral Reform. State and Local Government Review.31:147-152.

Maddala, G. S. (1983). Limited Dependent and Qualitative Variables in Econometrics. Cambridge:Cambridge University Press.

Mangat, N. S. & Singh, R. (1990). An alternative randomised response procedure. Biometrica77(2): 439–442.

Mangat, N. S. (1994). Use of a Modified Randomization Device in Warner's Model, Journal of the Indian Society of Statistics & Operational Research, 16, (1995) 65-69.

Martin, O. (2009). Assessing Sensitive Attributes Using the Randomized Response Technique: Evidence for the Importance of Response Symmetry, Journal of Educational and Behavioral Statistics June 2009, Vol. 34, No. 2, pp. 267–287 DOI: 10.3102/1076998609332747

Nayak, T. K. (1994) Randomized Response Surveys for Estimating a Proportion, Communications in Statistics-Theory and Methods, 23(11), (1994), 3303-3321. Orvieto, Italy, pp. 341–348.

Perri, Pier Francesco; Pelle, Elvira and Stranges, Manuela (2015): Estimating Induced Abortion and Foreign Irregular Presence Using the Randomized Response Crossed Model. Social Indicators Research. DOI 10.1007/s11205-015-1136-x

Quatember, A. (2013). A Standardization of Randomized Response Strategies. Survey Methodology, 35(2), 143-152.

Rasinski, K. A., Willis, G. B., Baldwin, A. K., Yeh, W. & Lee, L. (1999). Statistical response model. Proceedings of the Social Statistics Section, ASA: 65–72. .

Ruenda B. & Perri, P.F (2018). Randomized Response estimation in multiple frame surveys. International Journal of computer Mathematics, Volume 97, issue 1-2.

Sidhu Sukhjinder Singh, Bansal Mohan Lal, Kim Jong-Min, Singh Sarjinder (2009): unrelated question model in sensitive multi-character surveys. Communications of the Korean Statistical Society, volume 16, No.1, 169-183.

2019 Report on violence against children survey by UNICEF.

Walter, F. & Preisedover P. (2013). Asking sensitive questions; an evaluation of randomized response technique versus direct questioning. Sociological methods and research, 42,324-353,(1305).

Warner, S. L. (1965). Randomised response: A survey technique for eliminating evasive answer Warner, S. L. (1971). The linear randomised response model. Journal of the American Statistical Association 66 (336): 884–888.

Zawar Hussain , Javid Shabbir& Muhammad Riaz (2010) Bayesian Estimation Using Warner's Randomized Response Model through

Simple and Mixture Prior Distributions, Communications in Statistics—Simulation and Computation®, 40:1, 147-164.

APPENDICES

Appendix 1: Secondary school Students questionnaire

1. What is your gender?

(a) Male ( )

(b) Female ( )

3.What is your class?..........................

4.What is the type of your admission number?

(a) Even number ( )

(b) Odd number ( )

5. If your admission number is an even number, please reply “yes” to the following question independently of its content, if however, your admission number is odd number, then please answer truthfully.

Have you ever cheated in examination? Yes ( ) No ( )

6. If your admission number is an odd number, then please reply “no” to the following question independently of its content. If, however your admission number is an even number, then please answer truthfully."

Have you eve cheated in examination? Yes ( ) No ( )