International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2019; 4(4): 50-55

www.mathsjournal.com
Received: 22-05-2019
Accepted: 24-06-2019

S Mbala
University of Nairobi, P.O Box 30197-00100, Nairobi, Kenya

MM Manene
University of Nairobi, P.O Box 30197-00100, Nairobi, Kenya

JAM Ottieno
University of Nairobi, P.O Box 30197-00100, Nairobi, Kenya

Correspondence
S Mbala
University of Nairobi, P.O Box 30197-00100, Nairobi, Kenya

Symmetric truth detection model: A randomized response approach

S Mbala, MM Manene and JAM Ottieno

Abstract
When collecting sensitive information on abortion, drug addiction, examination dishonesty and tax evasion among others, many researchers use direct questioning which may not yield valid data. This is because respondents fear embarrassment and victimization. In this study we have formulated a Symmetric Truth Detection Model which uses two randomization devises to protect the privacy of respondents leading to a more honest response. This model is more efficient than the earlier models namely the Asymmetric Truth detection Models.

Keywords: Randomized response, symmetric, asymmetric, sensitive questions, sensitive attribute, randomization devise, truth detection

1. Introduction
Self-report is one of the most frequently used data collection techniques in research. However, people do not always tell the truth when being asked to answer sensitive questions (Martin, et.al. 2009) [1]. 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.

2. Literature Review
Warner (1965) [2] developed an interviewing procedure designed to reduce errors caused by non-response and untruthful 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.
A lot of improvement has been done on Warner model. Mangat (1994) [3] proposed a technique in which some basic demographic questions together with questions unrelated to the current study and the sensitive question are included in the questionnaire aimed at increasing the privacy of the respondents. Eichhorn and Hayre (2003) [4] extended the model developed by Mangat (1994) [3] 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. Zawar et.al. (2010) [5] 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) [6] 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.

All the models discussed above used one randomization devise and were characterized by large variances and a high level of non response. However no effort has been made to estimate the proportion of sensitive attribute using two randomization devises. This led to the formulation of symmetric TDM which uses two randomization devises to estimate the stigmatizing attribute.

3. Data and Methodology

To estimate the variance of the sensitive attribute, data simulation was used on both the Asymmetric and Symmetric TDM and relative efficiency was computed to ascertain the model which has the lowest variance.

3.1 Asymmetric Truth Detection Models

The Asymmetric TDMs are formulated by use of one randomization devise to collect sensitive attribute (Bo Yu, 2015) [6]. In these models, a single randomization devise such as a coin, a card, a spinner or any other devise is used and a "no" response identifies an interviewee as not holding the sensitive characteristic. In these models, 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*. In Asymmetric models, with population categories A and A^c

a box with two types of cards labeled A and A^c (in proportion *p:1- p*) is used as the randomization device. A respondent draws a card at random and uses it to select one of the two statements given below;

(i) I belong to group A

(ii) I do not belong to group A

Where group A represents those who have the sensitive attribute.

The respondents then have the option of responding 'yes' or 'no' according to whether or not they belong to the group of sensitive attribute. The Asymmetric model is formulated using the following notations;

Let λ be the observed proportion of ''yes" answers in the sample, *p* be the probability that arespondent 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.

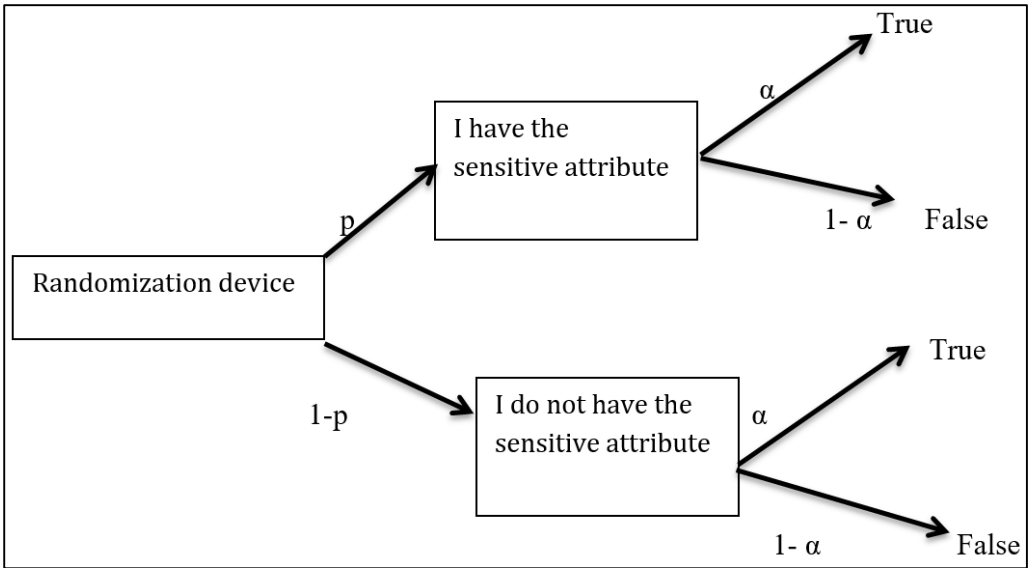
Proposition 2.1

The estimator for the sensitive attribute α for the asymmetric TDM is estimated as;

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

Proof

Proposition 2.1 can be proofed using the probability tree diagram below; True



In the tree diagram above, *p* is 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. 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). This can be expressed as below;

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

Which can be expanded as;

$\lambda = \alpha\,(2p - 1) + 1 - p.$ (2)

That is

$$\alpha = \frac{\lambda+(1-p)}{2p-1} \quad (3)$$

The proportion of sensitive attribute α is thus estimated as

$$\hat{\alpha} = \frac{\hat{\lambda}+(1-p)}{2p-1} \quad (4)$$

Swhere $\hat{\lambda}$ is an unbiased estimator for the total proportion of "yes" responses λ
This completes the proof.

Proposition 2.2
The variance of $\hat{\alpha}$ for asymmetric TDM is ;

$$Var\,(\hat{\alpha}) = \frac{\alpha(1-\alpha)}{n} + \frac{p(1-p)}{n(2p-1)^2} \quad (5)$$

Proof

Using equation (4) above;

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

$$= Var\left(\frac{\hat{\lambda}}{2p-1}\right) + Var\left(\frac{1-p}{2p-1}\right). \quad (6)$$

But, the variance of a constant = 0 therefore,

$$Var\left(\frac{1-p}{2p-1}\right) = 0.$$

Equation (6) therefore simplifies to;

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

which is equivalent to;

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

Using equation (2),

$1-\lambda = \alpha\,(1 - 2p) + p$

Hence;

$$Var(\hat{\alpha}) = \frac{[\alpha\,(2p - 1) + 1 - p][\alpha\,(1 - 2p) + p]}{n(2p-1)^2}. \quad (7)$$

Equation (7) can be expanded to;

$$Var(\hat{\alpha}) = \frac{(4p^2\alpha - 4p\alpha + \alpha - 4p^2\alpha^2 - 4\alpha^2 p + \alpha^2 + p - p^2)}{n(2p-1)^2}.$$

This simplifies to;

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

This completes the proof.

3.2 Symmetric Truth Detection Model

In this paper we have developed a new model which we have called Symmetric TDM. The Symmetric TDM was developed to overcome the problems associated with the Asymmetric truth detection model developed by Warner (1965)[2] and improved by BoYu, *et.al* (1015)[6]. These problems are;

1. Low privacy level
2. Large variance leading to invalid data.
3. Non response.

This model uses two randomization devises namely; D_1 and D_2 thus improving the privacy level as well as reducing variance and non response. Simple random sampling with replacement was used to select a sample of size n to estimate the proportion (α) of individuals who possess the sensitive attribute. Let a and b be any two positive real number, q be the probability of using the first randomization devise D_1 and 1-$q$be the probability of using the second randomization devise such that;

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

In this model the respondents are first presented with two randomization devises D_1 and D_2. They are then allowed to select one device freely. After which they are presented with two statements as shown below:

i. I belong to group *A*

ii. I do not belong to group *A*,

where group *A* represents those who have the sensitive attribute.

In the first randomization device D_1 the selection of statement (i) or (ii) is done with probabilities p_1 and 1-p_1 respectively. In the second randomization device D_2 the selection is done with probabilities p_2 and 1-p_2 for statements (i) and (ii) respectively. The probability of "yes" response (λ) is given by;

$$\lambda = \frac{a}{a+b}\{p_1\,\alpha + (1 - p_1)(1 - \alpha)\} + \frac{b}{a+b}\{p_2\,\alpha + (1 - p_2)(1 - \alpha)\} \quad (8)$$

Theorem 3.1

The unbiased estimator of α is given by;

$$\hat{\alpha} = \frac{\hat{\lambda}(a+b) - p_1 b - p_2 a}{(2p_1 - 1)(a-b)} \quad (9)$$

Proof

Using equation (8), the proportion of "yes" responses is given by:

$$\lambda = \frac{a\{p_1\,\alpha + (1 - p_1)(1 - \alpha)\} + b\{p_2\,\alpha + (1 - p_2)(1 - \alpha)\}}{a+b} \quad (10)$$

After simplification, equation (10) reduces to;

$$\lambda = \frac{\alpha\{(2p_1 - 1)(a - b)\} + \{p_1 b + p_2 a\}}{a+b}. \quad (11)$$

This implies that,

$$\alpha = \frac{\lambda(a+b) - p_1 b - p_2 a}{\{(2p_1 - 1)(a-b)\}} \quad (12)$$

The unbiased estimator of α is thus given by;

$$\hat{\alpha} = \frac{\hat{\lambda}(a+b) - p_1 b - p_2 a}{\{(2p_1 - 1)(a-b)\}} \quad (13)$$

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

This completes the proof.

Theorem 3.2

The variance of $\hat{\lambda}$ is given by;

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

Proof

Using equation (13) above;

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

This simplifies to;

$$Var(\hat{\alpha}) = \text{Var}\left(\frac{\hat{\lambda}(a+b)-p_1\,b-p_2\,a}{(2p_1-1)(a-b)}\right) - \left(\frac{p_1\,b-p_2\,a}{(2p_1-1)(a-b)}\right) \quad (14)$$

But the variance of a constant is 0, therefore;

$$Var(\hat{\alpha}) = \text{Var}\left(\frac{-p_1\,b+a}{(2p_1-1)(a-b)}\right) = 0.$$

Equation (14) therefore reduces to;

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

$$= \left(\frac{Var(\hat{\lambda})(a+b)^2}{(2p_1-1)^2(a-b)^2}\right) \quad (16)$$

But Var(λ) = $\left(\frac{\lambda(1-\lambda)}{n}\right)$ (17)

Substituting equation (17) in equation (16), we get;

$$Var(\hat{\alpha}) = \left(\frac{\alpha\{(2p_1-1)(a-b)\}+\{p_1\,b+p_2a\}\{(a+b)-(\alpha\{2p_1-1)(a-b)\}+\{p_1\,b+p_2a\}(a+b)^2}{n(2p_1-1)^2(a-b)^2(a+b)^2}\right) \quad (18)$$

This reduces to;

$$Var(\hat{\alpha}) = \frac{2a\alpha(1-p_1-p_2)}{n(2p_1-1)^2(a-b)^2(a+b)^2} + \frac{(p_2a-p_1b)((p_2a-p_1b)}{n(2p_1-1)^2(a-b)^2(a+b)^2} + \frac{\alpha(1-\alpha)}{n} \quad (19)$$

By setting $p_1 = 1 - p_2$

Equation (19) reduces to;

$$\text{Var}(\hat{\alpha}) = \frac{(p_2a+p_1b)(p_1\,a+p_2\,b)}{n(2p_1-1)^2(a-b)^2(a+b)^2} + \frac{\alpha(1-\alpha)}{n}. \quad (20)$$

This completes the proof.

4. Results and Discussion

The relative efficiency (RE) of the Symmetric TDM with respect to Asymmetric TDM is obtained by dividing the variance for Asymmetric TDM by the variance of Symmetric TDM after obtaining data using simulation method. The proposed model (Symmetric TDM) will be more efficient than Asymmetric TDM if;

$$\text{RE} = \frac{\text{variance for Asymmetric TDM}}{\text{variance for symmetric TDM}} > 1$$

This was shown in Table 1 below.

We have done the comparison of the variance of the Asymmetric TDM and Symmetric TDM by setting n=10, p=0.3, $p_1 = 0.1$,$\alpha = 0.7$, b = 2, a – b > 0 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 1: Relative efficiency

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

The results in Table 1 shows that, the Symmetric Truth Detection Model has a less variance compared to Asymmetric TDM and therefore more efficient than the Asymmetric Truth Detection Model since the RE > 1. It can be observed that efficiency increases with increase in the difference between *a* and *b*. This is because a bigger difference would increase the denominator thus decreasing the variance of Symmetric TDM and consequently increasing the efficiency.

5. Conclusion

Based on the results in table 1, the Symmetric Truth Detection Model has less variance than the Asymmetric Truth Detection Model and therefore more efficient for collecting data on sensitive information.

6. References

1. Martin O *et al*. Assessing Sensitive Attributes Using the Randomized Response Technique: Evidence for the Importance of Response Symmetry Journal of Educational and Behavioral Statistics. Journal of Educational and Behavioral Statistics. 2009; 34(2):267–287.
2. Warner SL. Randomized response: A Survey Technique for Eliminating Evasive Answer Bias, Journal of the American Statistical Association. 1965; 60:63-69.
3. Mangat NS, Singh R. An alternative randomized response procedure. Biometrika. 1994; 77:439-442.
4. Eichhorn BH, Hayre LS. Scrambled Randomized Response Methods for Obtaining Sensitive Quantitative Data. J of Statistical planning and Inference. 2003; 7:307–316.
5. Zawar Hussain, Javid Shabbir, Muhammad Riaz. Bayesian Estimation Using Warner's Randomized Response Model through Simple and Mixture Prior Distributions, Communications in Statistics—Simulation and Computation®. 2010; 40(1):147-164.
6. Bo Yu, Zongda J, Jiayong T, Ge G. Estimation of sensitive information by randomized response data in successive sampling. 2015; 4(13):29-42.