Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 70(3) 2016 197–203

An Improved Two Stage Optional RRT Model

Neeraj Tiwari and Prachi Mehta
Kumaun University, S.S.J. Campus, Almora

Received 12 July 2016; Accepted 23 July 2016

SUMMARY

Randomized Response Technique (RRT) is an effective survey method for collecting data on sensitive issues such as drug uses, tax evasion and induced abortion etc., while trying to maintain respondent's anonymity. Originally, RRT has been proposed by Warner (1965). Gupta (2001) and Sihm and Gupta (2014) considered two unknown parameters π and ω in their model and estimated them taking two independent samples. In this paper, we have proposed an improved methodology for RRT, in which the sensitivity level (ω) is considered to be known and the RR technique was applied only for those respondents who considered the particular question to be sensitive. This makes the procedure simpler and more efficient compared to the procedure of Gupta (2001) and Sihm and Gupta (2014). It has been theoretically established that the variance of the proposed estimator is less than the variance of the estimators suggested by Warner (1965) and Mangat and Singh (1990) under specific conditions. Some numerical examples have also been considered to demonstrate the utility of the proposed procedure over the existing RRT procedures.

Keywords: Optional randomized response model, Randomization device, Sensitivity level, Scrambling variable.

1. INTRODUCTION

Gaining valid answers to survey questions with sensitive behavior is difficult. Most of the people avoid giving answers of such sensitive questions. To overcome this difficulty various techniques have been developed which attempt to maintain the respondent's privacy so that more frank and honest answers can be extracted from the respondents. Randomized Response Technique (RRT) Warner (1965), Bogus Pipeline Technique (BPT) Jones and Sigall (1971) and Unmatched Count Technique (UCT) Raghavarao and Federer (1979) are such techniques to avoid difficulties of non response. Here we restrict ourselves to the Randomized Response Technique (RRT) proposed by Warner (1965).

To collect sensitive information from the respondents, a simple random sample of size 'n' is drawn with replacement from a population of 'N' individuals. Warner (1965) suggested a randomization device, in which each respondent chooses one of the two questions–

1. "Do you belong to A?",

and 2. "Do you not belong to A?",

with probabilities 'p' and '$(1–p)$' respectively. Here 'A' denotes the sensitive group. Interviewer gets only the answer 'yes' or 'no'. After getting the 'yes' or 'no' responses, interviewer estimates the value of required population proportion (π) using the estimator suggested by Warner.

In the model proposed by Warner (1965), both the questions were related to the same sensitive group. To protect the respondent's privacy, it is required that the two questions be unrelated so that more truthful answer can be extracted. Greenberg *et al.* (1969) introduced an unrelated question technique in which instead of the second question in the model suggested by the Warner (1965), respondents were asked to answer the question, which was unrelated to the sensitive character, such as – 'Are you born in the month of July?'

In order to increase the efficiency of the estimator and for getting more cooperation from

Corresponding author: Neeraj Tiwari
E-mail address: kumarn_amo@yahoo.com

respondents, Mangat and Singh (1990) proposed 'Two Stage Randomized Response (RR) Model'. In this model each respondent is provided with two randomization device. First randomization device consists of two statements-(1) 'I belong to the sensitive group' and (2) 'go to the second randomization device'. The second randomization device followed the procedure given by Warner (1965).

Hong *et al.* (1994) presented a stratified RRT, in which the whole population was divided into strata and then a proportional sample was selected from each stratum. By this method a researcher obtained a better sample to collect sensitive information.

An Optional Randomized Response Model was first introduced by Gupta (2001). He used the idea that a question may be sensitive to one person but not to the another. In this optional model, the respondent was free to choose how to answer the question. The respondent gives a scrambled answer if he/she seems the survey question is sensitive and gives a true response if he/she feels the survey question is non-sensitive. So there is a choice or an option for the respondents to give their answer.

It may be costly to take proportional sample from each stratum as given in the model presented by Hong *et al.* (1994). To reform this problem, Kim and Warde (2004) suggested a stratified RRT using an optimal allocation which was more efficient than that using a proportional allocation.

Odumade and Singh (2009) proposed RR model based on two decks of cards. By adjusting the proportion of cards in the two decks of cards, their model becomes more efficient than the models suggested by Warner (1965), Mangat and Singh (1990) and Mangat (1994).

Sihm and Gupta (2014) suggested Two-Stage Binary Optional Randomized Response Model which was based on optional RR model given by Gupta (2001). The focus of this model was on estimating π (class possessing the sensitive character) and ω (sensitivity level of the survey question). Their method gives better results than optional RRT given by Gupta (2001).

The model suggested by Gupta (2001) and Sihm and Gupta (2014) had the drawback that ω and π were unknown quantities, which were estimated using two samples from the population. This made their procedure practically tedious. In this paper, we have separated the group that does not feel that the particular question is sensitive from the sample and suggested to receive the direct response from them. Using this procedure, we can get the value of the sensitivity level (ω). This makes the procedure practically easier because ω is assumed to be known quantity and as such π can be estimated using only one sample. Under specific conditions the variance of the proposed estimator comes out to be lesser than the variance of the estimators suggested by Warner (1965) and Mangat and Singh (1990). The utility of the proposed procedure has been demonstrated with the help of some numerical examples.

In section 2, a brief description of some previous RRT models has been presented. The proposed model has been described in section 3. In section 4, we have compared variance of the proposed estimator with the variance of the estimators presented in section 2. In section 5, some numerical examples have been considered to demonstrate the utility of the proposed procedure. The findings of the paper have been discussed in section 6.

2. A BRIEF DESCRIPTION OF SOME PREVIOUS RRT MODELS

2.1 Warner (1965)

By using a randomization device, Warner gave an option to the respondent for choosing one of the two questions:

1. "Do you possess the sensitive character A?"

and 2. "Do you not possess the sensitive character A?"

with probability 'p' and '$(1–p)$' respectively. By using the randomization device respondent give a 'yes' or 'no' answer either to the sensitive question or to its negative depending on the outcome of that randomization device. Since the interviewer would not know the outcome of the device and would get only 'yes' or 'no' answer, the privacy of the respondent could be maintained.

Let P_y be the probability of 'yes' response and π be the class possessing the sensitive character, then we have

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

Solving for π, we get

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

The Warner's estimate of π is

$$\hat{\pi}_w = \frac{\hat{P}_y - (1-p)}{2p-1}\text{, and} \tag{1}$$

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

2.2 Mangat and Singh (1990)

In this model, to have more truthful answer, one more randomization device was added into the model given by Warner (1965). This procedure has been completed in two stages and is known as Two-Stage RR model. In the first stage, first randomization device has two options-(1) "Do you possess the sensitive character A?" and (2) "Go to the Warner's randomization device" with the probability 'T' and '(1-T)' respectively. The entire process remains unobserved by the researcher, so the privacy of the respondents has been maintained. In this method,

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

Solving for π, $$\hat{\pi}_m = \frac{\hat{P}_y - (1-p)(1-T)}{(2p-1)+2T(1-p)}, \tag{3}$$

where P_y is the proportion of 'yes' responses. The variance of the estimate was given by,

$$\text{Var}\left(\hat{\pi}_m\right) = \frac{1}{\{(2p-1)+2T(1-p)\}^2}\left\{\frac{P_y(1-P_y)}{n}\right\} \tag{4}$$

2.3 Gupta (2001)

This model is based on the assumption that some population proportion does not feel a particular question to be sensitive and as such they can give straight answer to that question, if they get the option to answer truthfully. Accordingly, they are provided an opportunity to answer truthfully and rest of the people who feel the question is sensitive follow the method given by Warner (1965). The entire process takes place without the knowledge of the interviewer and maintained the privacy of the respondents.

If ω be the sensitivity level and P_y be the probability of 'yes' response, then

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

Or $P_y - \pi = (p-1)(2\pi-1)\omega$

In this model π and ω are two unknown parameters, hence two samples were needed to estimate π and ω with the sample sizes n_1 and n_2 respectively. Hence we have

$$P_{y_1} - \pi = (p_1 - 1)(2\pi - 1)\omega$$

and $P_{y_2} - \pi = (p_2 - 1)(2\pi - 1)\omega$

Solving the above equation for π,

$$\hat{\pi}_g = \frac{\lambda \hat{P}_{y_2} - \hat{P}_{y_1}}{\lambda - 1} \tag{5}$$

where $\lambda = \dfrac{p_1 - 1}{p_2 - 1}$

The variance of the estimate was given by,

$$\text{Var}\left(\hat{\pi}_g\right) = \frac{1}{(\lambda-1)^2}\left\{\lambda^2\frac{P_{y_2}(1-P_{y_2})}{n_2} + \frac{P_{y_1}(1-P_{y_1})}{n_1}\right\} \tag{6}$$

And estimator for sensitivity level (ω) was given by

$$\hat{\omega}_g = \frac{\hat{P}_{y_1} - \hat{P}_{y_2}}{2\hat{P}_{y_1}(1-p_2) - 2\hat{P}_{y_2}(1-p_1) - (p_1 - p_2)} \tag{7}$$

2.4 Sihm & Gupta (2014)

In this model, in the first stage, first randomization device has two options–(1) "Do you possess the sensitive character A?" and (2) "Go to the second randomization device" with the probability T and (1-T) respectively. The second randomization device follows the method proposed by Gupta (2001). Then the probability of 'yes' response is

$$P_y = T\pi + (1-T)\left[(1-\omega)\pi + \omega\{\pi p + (1-\pi)(1-p)\}\right]$$

Like Gupta model, two independent samples were needed for estimating π and ω, with sample sizes n_1 and n_2 respectively. Hence

$$\hat{\pi}_p = \frac{\lambda \hat{P}_{y_2} - \hat{P}_{y_1}}{\lambda - 1} \tag{8}$$

where, $\lambda = \dfrac{p_1 - 1}{p_2 - 1}$

The variance of the estimate was given by,

$$\mathrm{Var}\left(\hat{\pi}_p\right) = \frac{1}{(\lambda-1)^2} \left\{\lambda^2 \frac{P_{y_2}\left(1-P_{y_2}\right)}{n_2} + \frac{P_{y_1}\left(1-P_{y_1}\right)}{n_1}\right\} \tag{9}$$

And estimator for sensitivity level (ω) was given by

$$\hat{\omega}_p = \left\{\frac{P_{y_1} - P_{y_2}}{(1-T)(p_1 - p_2)}\right\} \frac{1}{(2\pi - 1)} \tag{10}$$

$$(p_1 \neq p_2)$$

3. THE PROPOSED IMPROVED TWO STAGE OPTIONAL RRT MODEL

It may be possible that any survey question which is sensitive for some people may not be sensitive for other people. For example a question such as "Are you a smoker?" may not be sensitive for some people and they can give an honest answer to the interviewer.

The population proportion, who consider the question is sensitive is known as sensitivity level and denoted by ω. Gupta (2001) used this idea of sensitivity level in his optional model and considered ω as unknown value. Our main concern is that if the respondent feels that the question is not sensitive for him/her and can give a straight answer of it, then we can separate this population from the remaining population which feels that the question is sensitive and can get the direct response from them. By this way we can directly get the value of ω and need not to estimate it as suggested by Gupta (2001) and Sihm and Gupta (2014). Rest of the people who feel the question is sensitive (ω) give the scrambled response. Since in the proposed model the value of ω is already known to us, there is no need to take two samples to estimate the required population proportion. So it would be practically easier for the interviewer to collect sensitive information using the proposed method.

In the proposed model, in first step, we separate the respondents, who feel the survey question is non-sensitive and get the true response from them. In the second step, rest of the respondents follows the procedure given by Mangat and Singh (1990) in which two randomization device were used. The first randomization device has two options:

1. "Do you belong to the sensitive group?" and
2. "Go to the second randomization device"

with probabilities T and $(1–T)$ respectively. The second randomization device is same as the Warner's randomization device. The privacy of the respondents (who feels the question is sensitive) can be maintained because the procedure in the second step remains unobserved by the researcher. There is no need to maintain the privacy of the respondents who feel that the question is non-sensitive for them.

Let π be the population proportion related to the sensitive group. The probabilities of the two questions used in Warner's device are p and $(1–p)$ respectively. If P_y be the probability that a respondent gives 'yes' answer, then we have

$$P_y = (1-\omega)\pi + \omega[T\pi + (1-T)\{\pi p + (1-\pi)(1-p)\}]$$

$$\text{Or } P_y = \pi(1 - 2\omega + 2\omega T + 2\omega p - 2\omega Tp) + \omega(1 - p - T + Tp)$$

Solving the above equation for π,

$$\hat{\pi}_n = \frac{\hat{P}_y - \omega(1-p)(1-T)}{\{1 - 2\omega(1-p)(1-T)\}} \tag{11}$$

where $\overset{\wedge}{\pi}_n$ is the estimator of the proposed model. The variance of the proposed estimator is given by,

$$\mathrm{Var}\left(\overset{\wedge}{\pi}_p\right) = \frac{1}{\{1-2\omega(1-p)(1-T)\}^2}\left\{\frac{P_y(1-P_y)}{n}\right\} \quad (12)$$

4. EFFICIENCY COMPARISON

In this section, we compare the efficiency of the proposed estimator in comparison to the existing estimators.

4.1 Comparison of the Proposed Estimator with the Estimator Given by Warner (1965)

Variance of the estimator proposed by Warner (1965):

$$\mathrm{Var}\left(\overset{\wedge}{\pi}_w\right) = \frac{1}{(2p-1)^2}\left\{\frac{P_y(1-P_y)}{n}\right\}$$

Variance of the proposed estimator:

$$\mathrm{Var}\left(\overset{\wedge}{\pi}_n\right) = \frac{1}{\{1-2\omega(1-p)(1-T)\}^2}\left\{\frac{P_y(1-P_y)}{n}\right\}$$

$$\mathrm{Var}\left(\overset{\wedge}{\pi}_w\right) - \mathrm{Var}\left(\overset{\wedge}{\pi}_n\right) = \frac{P_y(1-P_y)}{n}$$

$$\left\{\frac{1}{(2p-1)^2} - \frac{1}{\{1-2\omega(1-p)(1-T)\}^2}\right\}$$

$$\mathrm{Var}\left(\overset{\wedge}{\pi}_w\right) - \mathrm{Var}\left(\overset{\wedge}{\pi}_n\right) > 0$$

If $\{1-2\omega(1-p)(1-T)\}^2 - (2p-1)^2 > 0$

i.e. if $\omega^2(1-p)(1-T)^2 - \omega(1-T) + p > 0$

4.2 Comparison of the Proposed Estimator with the Estimator Given by Mangat and Singh (1990)

Variance of the estimator given by Mangat and Singh (1990):

$$\mathrm{Var}\left(\overset{\wedge}{\pi}_m\right) = \frac{1}{\{(2p-1)+2T(1-p)]^2}\left\{\frac{P_y(1-P_y)}{n}\right\}$$

Variance of the proposed estimator:

$$\mathrm{Var}\left(\overset{\wedge}{\pi}_n\right) = \frac{1}{\{1-2\omega(1-p)(1-T)\}^2}\left\{\frac{P_y(1-P_y)}{n}\right\}$$

$$\mathrm{Var}\left(\overset{\wedge}{\pi}_m\right) - \mathrm{Var}\left(\overset{\wedge}{\pi}_n\right) = \frac{P_y(1-P_y)}{n}$$

$$\left[\frac{1}{\{(2p-1)+2T(1-p)\}^2} - \frac{1}{\{1-2\omega(1-p)(1-T)\}^2}\right]$$

$$\mathrm{Var}\left(\overset{\wedge}{\pi}_m\right) - \mathrm{Var}\left(\overset{\wedge}{\pi}_n\right) > 0$$

If $\{1-2\omega(1-p)(1-T)\}^2 - \{(2p-1)+2T(1-p)\}^2 > 0$

i.e. if $0 < \omega < 1$

i.e. if ω lies between 0 and 1

Since under the given conditions, variance of the proposed estimator $\left\{\mathrm{Var}\left(\overset{\wedge}{\pi}_n\right)\right\}$ is less than the variance of the estimators given by Warner (1965) and Mangat and Singh (1990), we can conclude that the proposed estimator $\overset{\wedge}{\pi}_n$ is more efficient than the estimators given by Warner (1965) and Mangat and Singh (1990).

5. NUMERICAL EXAMPLES

In this section we have considered some numerical examples to demonstrate the utility of the proposed estimator.

Example 1. Let for the sample size 35, number of yes responses are 20 and the sensitivity level is 25. Let the Warner's randomization device be a dice in which respondent chooses first question if the dice comes with number 1 and 2 otherwise respondent goes for second question. Let the other randomization device be a coin. Then we have

$$n = 35,\ P_y = \frac{20}{35},\ \omega = \frac{25}{35},\ (1-\omega) = \frac{10}{35},$$

$$p = \frac{2}{6},\ (1-p) = \frac{4}{6},\ T = \frac{1}{2},\ (1-T) = \frac{1}{2}$$

Then $\mathrm{Var}\left(\overset{\wedge}{\pi}_w\right) = 0.44$

$\mathrm{Var}\left(\overset{\wedge}{\pi}_m\right) = 0.44$

$$\text{Var}\left(\hat{\pi}_n\right) = 0.03$$

In this example

$$\text{Var}\left(\hat{\pi}_n\right) < \text{Var}\left(\hat{\pi}_w\right)$$

and $\text{Var}\left(\hat{\pi}_n\right) < \text{Var}\left(\hat{\pi}_m\right)$

Example 2. Let for the sample size 20, number of yes responses are 12 and the sensitivity level is 15. Let the Warner's randomization device be a dice in which respondent chooses first question if the dice comes with number 1 otherwise respondent goes for second question. Let the other randomization device be a coin. Then we have

$$n = 20,\ P_y = \frac{12}{20},\ \omega = \frac{15}{20},\ (1-\omega) = \frac{5}{20},$$

$$p = \frac{1}{6},\ (1-p) = \frac{5}{6},\ T = \frac{1}{2},\ (1-T) = \frac{1}{2}$$

Then $\text{Var}\left(\hat{\pi}_w\right) = 0.027$

$$\text{Var}\left(\hat{\pi}_m\right) = 0.432$$

$$\text{Var}\left(\hat{\pi}_n\right) = 0.085$$

In this example

$$\text{Var}\left(\hat{\pi}_n\right) > \text{Var}\left(\hat{\pi}_w\right)$$

because $\omega^2(1-p)(1-T)^2 - \omega(1-T) + p < 0$

and $\text{Var}\left(\hat{\pi}_n\right) < \text{Var}\left(\hat{\pi}_m\right)$

Example. 3. Let for the sample size 50, number of yes responses are 20 and the sensitivity level is 35. Let the Warner's randomization device be a bag containing 22 red and 38 green balls using which respondent chooses first question if red ball comes out from the bag and the respondent goes for second question if he/she gets green ball. Let the other randomization device be a dice in which $T = 4/6$. Then we have

$$n = 50,\ P_y = \frac{20}{50},\ \omega = \frac{35}{50},\ (1-\omega) = \frac{15}{50},$$

$$p = \frac{22}{60},\ (1-p) = \frac{38}{60},\ T = \frac{4}{6},\ (1-T) = \frac{2}{6}$$

Then $\text{Var}\left(\hat{\pi}_w\right) = 0.067$

$$\text{Var}\left(\hat{\pi}_m\right) = 0.014$$

$$\text{Var}\left(\hat{\pi}_n\right) = 0.005$$

In this example

$$\text{Var}\left(\hat{\pi}_n\right) < \text{Var}\left(\hat{\pi}_w\right)$$

and $\text{Var}\left(\hat{\pi}_n\right) < \text{Var}\left(\hat{\pi}_m\right)$

6. DISCUSSION

In this article, we have proposed a model which may be considered as an improvement over the models given by Gupta (2001) and Sihm and Gupta (2014). While extracting information of sensitive nature the existing models had considered the sensitivity level (ω) as an unknown quantity. In the proposed model ω is considered as a known quantity under the assumption that if the respondent does not feel the survey question is sensitive and give an honest answer frankly, then it is not needed to conceal his/her identity and we can receive the true response from them without applying the randomization techniques. By this way ω is automatically known to the interviewer and RRT would be applied only for this group (ω). By using this procedure, it would be practically easier for the interviewer to collect relevant information from the respondents. The proposed model comes out to be more efficient than the models given by Gupta (2001) and Sihm and Gupta (2014) under certain conditions. We have also shown the utility of the proposed method over the existing RRT procedures with the help of some examples. The proposed procedure may be highly advantageous for collecting the sensitive information.

REFERENCES

Greenberg, B.G., Abul-Ela, As. L.A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question randomized response model- theoretical framework. J. *Amer. Statist. Assoc.*, **64(326)**, 520-539.

Gupta, S.N. (2001). Qualifying the sensitivity level of binary response personal interview survey questions. *J. Combi. Inf. Sys. Sci.,* **26(1-4),** 101-109.

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

Jones, E. and Sigall, H. (1971). The bogus pipeline: a new paradigm for measuring affect and attitude. *Psycho. Bull.,* **76(5),** 349-364.

Kim, J.M. and Warde, W.D. (2004). A stratified Warner's randomized response model. *J. Statist. Plann. Infer.,* **120(1-2),** 155-165.

Mangat, N.S. (1994). An improved randomized response strategy. *J. Roy. Statist. Soc.,* **B56**, 93-95.

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

Odumade, O. and Singh, S. (2009). Efficient use of two decks of cards in randomized response sampling. *Comm. Statist.-Theory Methods,* **34(4),** 439-446.

Raghavarao, D. and Federer, W.T. (1979). Block total response as an alternative to the randomized response method in surveys. *J. Roy. Statist. Soc.,* **B41(1)**, 40-45.

Sihm, J.S. and Gupta, S.N. (2014). A two stage binary optional randomized response model. *Comm. Statist.-Sim. Commu.*, **44(9)**, 2278-2296.

Warner, S.L. (1965). Randomized response : A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.*, **60(309)**, 63-69.