Statistics and Applications {**ISSN 2452-7395** (online)}
Volume 17, No. 2, 2019 (New Series), pp 47-54

Privacy Protection in Optional Randomized Response Surveys for Quantitative Characteristics

Dipika Patra and Sanghamitra Pal
West Bengal State University, Bārāsat

Received: August 04, 2019; Accepted: December 31, 2019

Abstract

A survey regarding sensitive or stigmatizing issues often bears a challenge as most respondents either deny answering direct queries or hide true response. Warner (1965) designed an ingenious device by a dint of a probabilistic procedure for estimating qualitative sensitive population proportion, called randomized response (RR) device and the novel technique is well known as randomized response technique (RRT). Greenberg et al. (1971) used the RR technique in quantitative attributes. An issue often raised with RRT is that some are more willing to answer directly rather than compulsory RR as the perception of sensitivity may not be same for all. Considering this fact, Chaudhuri and Mukerjee (1985) developed the optional randomized response technique (ORRT) which was restricted to Simple Random Sampling (SRS) design only. Later Chaudhuri and Saha (2005), Pal (2008) extended their work for unequal probability sampling. We discuss here about the privacy protection measure dealing with quantitative sensitive issues like alcohol consumption, earning through gambling, income etc. The literature is an extension of Chaudhuri and Christofides (2013) aiming at to develop how privacy is protected while applying ORRT into quantitative sensitive issues. In this paper we theoretically develop few well known RRTs for quantitative case in ORRT context first and unbiased estimators with its variance estimators are obtained. Protections of privacy of the proposed techniques are measured theoretically.

Key words: Protection of privacy, randomized response, sensitive issues, quantitative characteristics, unequal probability.

1. Introduction

Enumeration related to sensitive issues like alcohol consumption, drug addiction etc is usually impossible by direct survey method because the respondents may fear oppression if they disclose their actual status. Randomized response technique (RRT) refers to a widely used method for estimating population proportion or others which is related to a sensitive characteristic avoiding the direct queries. Warner (1965) developed the novel RRT technique. Erikson (1973) presented the estimation of total stigmatizing real variable like amount earns through gambling, alcohol consumption etc. Chaudhuri and Mukherjee (1985, 1988) illustrated optional randomized response technique (ORRT) while respondents were selected by SRS with replacement only. Later Arnab (2004) and Chaudhuri and Saha (2005) extended the theory in unequal probability design. Chaudhuri (2011), Chaudhuri and Christofides (2013) provide extensive developments in RRT, also in ORRT along with protection of privacy measures. For further references and recent developments, please refer to a monograph edited by Chaudhuri, Christofides, Rao (2016). Full ORRT and partial ORRT are

Corresponding Author: Sanghmitra Pal
Email: *mitrapal2013@gmail.com*

two classifications in ORRT. In the first one, overall sample of respondents are classified into two parts to gather direct response (DR)'s from one part and randomized response (RR) from another part. The final estimator was constructed by combining two linear unbiased estimators based on DR and RR separately. Arnab (2018) and many others had contributed in this full ORRT approach. In partial ORRT, DR option is offered along with a RR device to the respondents and respondents are requested to report the response directly if he/she does not feel the survey question sensitive otherwise it will be answered by using RR device without divulging the option so exercised.

The purpose of this paper is to extend some well-known quantitative RR technique in partial ORRT along with the study of their privacy protection capacity as the motivation of RRT is to gather information maintaining the respondents' confidentiality. Section 2 is designed for the extension of Chaudhuri's device I and device II (2011) and Eichhorn and Hayre (1983) in partial ORRT. In section 3, we briefly discuss the protection of privacy for different partial ORRT. In section 4, we provide the measures of jeopardy related to the proposed ORR models as discussed in section 2.

2. Proposed ORR Models for Quantitative Characteristics

Initially the basic purpose of RR was to gather reliable information on qualitative sensitive variables. Greenberg et al. (1971) first developed the RR technique for quantitative stigmatizing characteristics. Eriksson (1973) also extended the well-known unrelated question model in quantitative response concept. Taking the initial idea of masking true sensitive values by a random number from known distribution, Eichhorn and Hayre (1983) developed scrambled response model. Chaudhuri (2011) and Chaudhuri and Mukerjee (1988) in their books have mentioned two different randomized devices (Device I and Device II) along with the estimation of total amount and the variance relating to such sensitive issues. Following the idea of the above, we try to develop ORRT in quantitative measures which are as well as sensitive also.

Consider a finite population of units $U = (1, 2, 3, ..., N)$ and let Y be the quantitative stigmatizing variable having the values $y_1, y_2, ...y_i, ..., y_N$. A sample s of size n is chosen from the population according to a general sampling scheme P. The sampled persons are approached with a request to provide ORR responses for estimating the population total $Y = \sum_{i=1}^{N} y_i$ based on the sample s.

2.1. ORR using Eichhorn and Hayre (1983)

Pollock and Bek (1976) envisaged the data masking procedure to answer the sensitive question hiding his/her actual value by adding a random value from known distribution with the true value. The development of Eichhorn and Hayre (1983) method known as "Scrambled Response method" is actually in-depth analysis of Pollock and Bek (1976). In this part of this literature, we use the scramble response method to develop optional randomized response (ORR) model. Pal (2008) already worked on this by capturing two responses for each respondent giving an opportunity to report the second response as the earlier one with known probability p or by using other random variables with known probability $1 - p$. Our proposed ORR method is a modification on Pal (2008).

Let y_i be the true sensitive value of the i^{th} respondent $(i = 1,2,....,N)$. Let X denote a discrete random variable with known mean θ_1 and variance σ_1^2. Also, let Ψ be another discrete random variate independent to X, with known mean 0 and variance σ_2^2. Considering the fact that someone may wish to answer directly of the sensitive question, we give them a choice of direct response (y_i) or by randomized value (I_i) instead of compulsory RR. The procedure is known as ORRT as discussed in the introduction section. Here $I_i = \frac{y_i x_i}{\theta_1} + \psi_i$ while x_i and ψ_i are the values of the random variable X and Ψ respectively for i^{th} individual and $i = 1,2,..., N$.

Mathematically the response of i^{th} person may be written as,

$z_i = y_i$ with unknown probability $c_i \in [0,1]$

$= \frac{y_i x_i}{\theta_1} + \psi_i$ with unkonwn probability $1 - c_i$

Denoting E_R as expectation due to RR device and V_R as variance due to RR device, we get $E_R(z_i) = c_i y_i + (1 - c_i)\{\frac{y_i \theta_1}{\theta_1} + 0\} = y_i$ and

$V_R(z_i) = E_R(z_i^2) - E_R{}^2(z_i)$

$= c_i y_i^2 + (1 - c_i)\{\frac{y_i^2}{\theta_1^2}(\sigma_1^2 + \theta_1^2) + \sigma_2^2 + 2\frac{y_i \theta_1}{\theta_1}.0\} - y_i^2$

$= (1 - c_i)(\frac{y_i^2 \sigma_1^2}{\theta_1^2} + \sigma_2^2)$ which is unknown to us as y_i is unknown.

So, to estimate the variance, the whole process is repeated independently one more time to get another response z_i'. Technique of interpenetrating network of subsampling pioneered by Mahalanobis (1946) is used here to provide the final RR based estimator of y_i, which becomes $r_i = \frac{z_i + z_i'}{2}$ with variance estimator $v_i^* = \frac{1}{4}(z_i - z_i')^2$.

2.2. ORR using Chaudhuri' s device I

In ***device I***, the person labeled "i" is directed to give out his/her true response regarding the sensitive issues directly or by the offered two randomized devices. That process is repeated two times independently with the same RR device but with different RR parameters. In one RR device, first box contains T (>1) cards identical in shape, size, color and height bearing real numbers $a_1, a_2, a_3, ...a_T$ with mean $\mu_a = \frac{1}{T}\sum_i a_i = 1$ and the second box contains M (>1) identical cards with real numbers $b_1, b_2, b_3, ..., b_M$. In another RR device, T' cards $a_1', a_2', ...a_{T'}'$ bearing real numbers with mean 1 are placed in first box and M' cards $b_1', b_2', ...b_{M'}'$ bearing real numbers but the mean $\mu_{b'} = \frac{1}{M'}\sum_i b_i' \neq \mu_b = \frac{1}{M}\sum_i b_i$ are placed in the second box. The sampled person i is instructed to draw independently one card from each

box for both the RR devices and report the number $z_i = a_j y_i + b_k \ \ (j = 1,2,...,T \ \ k = 1,2,...,M)$ and $z_i' = a_j' y_i + b_k' \ \ (j = 1,2,...,T' \ \ k = 1,2,...M')$without disclosing the numbers drawn from the boxes and y_i is defined for the amount related to the sensitive quantitative variable Y.

In our proposed method, the optional randomized response for i^{th} person is

$z_i = y_i$ with unknown probability $c_i \in [0,1]$

$= a_j y_i + b_k$ with unknown probability $1 - c_i$

$z_i' = y_i$ with unknown probability c_i

$= a_j' y_i + b_k'$ with unknown probability $1 - c_i$

Writing, $E_R(z_i) = c_i y_i + (1 - c_i)(\mu_a y_i + u_b)$ and $E_R(z_i') = c_i y_i + (1 - c_i)(\mu_{a'} y_i + u_{b'})$. It follows $r_{1i} = \dfrac{\mu_{b'} z_i - \mu_b z_i'}{\mu_{b'} - \mu_b}$; $\mu_{b'} \neq \mu_b$ such that $E_R(r_{1i}) = y_i$.

To estimate the variance, the process is repeated independently one more time (as described in 2.1.) and the responses for the i^{th} person are (g_i, g_i') with corresponding estimator of y_i is $r_{2i} = \dfrac{\mu_{b'} g_i - \mu_b g_i'}{\mu_{b'} - \mu_b}$; $\mu_{b'} \neq \mu_b$. So the final RR based estimator of y_i is $r_i^* = \dfrac{r_{1i} + r_{2i}}{2}$ with the variance estimator $v_i^* = \dfrac{1}{4}(r_{1i} - r_{2i})^2$ such that $E_R(r_i^*) = y_i$.

2.3. ORR using Chaudhuri's device II

In ***device II***, a box with full of different kind of cards is given to the sampled person. The cards are marked as "corrected" with proportion k and others bearing with numbers $x_1, x_2, ..., x_M$ in proportion $q_1, q_2, ..., q_M$ such that $\sum_{i=1}^{M} q_i = 1 - k$. The sampled person is directed to draw a card randomly and report the true sensitive value $(i.e.\ y_i)$ if he gets a card marked as "corrected" otherwise report the number $x_j \ \ (j = 1,2,...,M)$ printed over the cards. The procedure is extended to ORRT by giving a choice of direct response or RR *device II* to the sampled person whichever he is willing too.

So the optional randomized response of i^{th} person is:-

$z_i = y_i$ with unknown probability $c_i \in [0,1]$

$= y_i$ with known proportion of cards (k) marked as "corrected" and unknown probability $(1 - c_i)$

$= x_j$ with unknown probability $q_j(1 - c_i)$

It follows that, $E_R(z_i) = c_i y_i + (1 - c_i)(k y_i + \sum_{j=1}^{M} q_j x_j)$. The whole process is repeated independently one more time with different set of cards in different proportion and the response is recorded as z_i'. Clearly, $E(z_i') = c_i y_i + (1 - c_i)(k' y_i + \sum_{j=1}^{M} q_j' x_j')$.

So, $E_R((1-k')z_i-(1-k)z_i')=(k-k')y_i+(1-c_i)[(1-k')\sum_{j=1}^{M}q_jx_j-(1-k)\sum_{j=1}^{M}q_j'x_j'].$

The sensitive attribute y_i is estimable if $(1-k')\sum_{j=1}^{M}q_jx_j-(1-k)\sum_{j=1}^{M}q_j'x_j'=0$.

i.e. $E_R[\frac{(1-k')z_i-(1-k)z_i'}{(k-k')}]=y_i$ if $\frac{1-k'}{1-k}=\frac{\sum_{j=1}^{M}q_j'x_j'}{\sum_{j=1}^{M}q_jx_j}$.

The final RR based estimator with the variance estimator can be obtained similarly as discussed in sections 2.1. and 2.2. in this article.

3. Privacy Protection Measures

The objective of performing RR survey is to produce a good estimator from statisticians' point of view for sensitive traits. As the respondents' actual state of nature is covered by RR device, it is necessary to know whether the procedure assures all the respondents that they could not definitely be classified in A or A^ci.e. protection of privacy measure. Undoubtedly, greater the protection; increase the participation but it has to be noted that no universally accepted measure is mentioned there. Privacy measure have been studied earlier by many researchers like Lanke (1975, 1976), Leysieffer and Warner (1976), Anderson (1975 a, b, c). Leysieffer and Warner (1976) suggested a jeopardy measure by quantifying the probability of an observation belonging to the sensitive trait A and its complement A^c, while giving his/ her response as R , termed as *revealing probabilities.* Lanke (1976) considered the quantity $g(A)=\max\{P(A\mid R=1),P(A\mid R=0)\}$ for comparison implying smaller the value of $g(A)$ is more protective than other. To evaluate how effectively the scrambling response model works, Eichhorn and Hayre (1983) proposed a privacy measure based on the ratio of the upper limit and lower limit of $100(1-\alpha)$confidence interval for the mean of the scrambling variable X . For a given α , larger the ratio implies greater the protection.

In order to evaluate how privacy is protected for quantitative sensitive variable is also investigated by Chaudhuri and Christofides (2013). Considering the prior unknown probability for the value y_i of i^{th} person as $L(y_i)=L_i$, by Bayes' theorem the posterior probability of y_i for the given value of z_i turns out to be,

$$L(y_i\mid z_i)=\frac{L_iP(z_i\mid y_i)}{P(z_i)} \tag{1}$$

where $P(z_i\mid y_i)$ denotes the conditional probability of reported RR value for the i^{th} person while true response is y_i. The degree of privacy protection measure is maximum, if the value of the measure approaches to one.

Taking a cue from the above approach, the idea of measuring protection of privacy for quantitative ORR model has been developed here.

4. Measures of Jeopardy

Suppose that y be the real stigmatizing quantitative variable for a set of finite population $U=(1,2,...,i,...,N)$ having values $y_1,y_2,....y_i,...y_N$ and the total be defined as $Y=\sum_{i=1}^{N}y_i$. In order

to estimate Y, a sample s is selected with probability $p(s)$ from the population U and respective ORR technique is performed to record their responses for further analysis.

4.1. Using Eichhorn and Hayre (1983) method

To check how well the response is protected in case of ORR survey while scrambled response method is used for RR value, responses are gathered by following the step by step guidance as described in the section 2.1. The conditional probability of the i^{th} person can be calculated by the following function

$$P(z_i \mid y_i) = c_i + (1-c_i)(\sum_{x_j} P(X = x_j)P(\Psi = z_i - \frac{y_i x_j}{\theta_1})),$$

as the respondent disclose the true response with probability c_i (at this point $z_i = y_i$) otherwise provide the randomized response with probability $1-c_i$ following Eichhorn and Hayre (1983) suggestion (as described briefly in section 2.2) which is equal to z_i if the randomized value of the variate X is x_i along with another variate (Ψ) value ψ_i.

Also, the probability that the i^{th} respondent gives the response z_i is defined as $P(z_i) = c_i + (1-c_i)(\sum_{x_j} P(X = x_j)P(\Psi = z_i - \frac{y_i x_j}{\theta_1}))$, this is exactly equal to the above given conditional probability.

So from equation (1), we get $L_i(y_i \mid z_i) = L_i$ *i.e.* posterior probability = prior probability. Clearly privacy is well protected for each individual by this method.

4.2. Using Chaudhuri's device I

For this model, the conditional probability of the given response z_i, while the actual response is y_i, denoted by $P(z_i \mid y_i)$, is evaluated as $P(z_i \mid y_i) = c_i + \frac{1}{TM}(1-c_i) = P(z_i)$. This indicates the posterior probability exactly equal to the prior probability L_i, *i.e.* the i^{th} respondent's privacy are well protected, as well as all the respondents.

4.3. Using Chaudhuri's device II

To check whether their response is well protected or not while device II is suggested for ORR survey, we calculate the posterior probability $L_i(y_i \mid z_i)$ by Bayes' theorem as defined in section 3 by following Chaudhuri and Christofides (2013). Here,

$$L_i(y_i \mid z_i) = \frac{L_i(c_i + k(1-c_i))}{L_i(c_i + k(1-c_i)) + (1-L_i)(1-c_i)(1-k)} = [1 + \frac{1-L_i}{L_i}\frac{\theta_i}{1-\theta_i}]^{-1}$$

considering $\theta_i = (1-c_i)(1-k)$.

Thus, $L_i(y_i \mid z_i)$ approaches to L_i if and only if $\theta_i \to \frac{1}{2}$. We can't say anything else as θ_i is unknown to us.

5. Concluding Remarks

Main purpose of this article is to demonstrate the accuracy level of privacy protection while studying quantitative and sensitive characteristics by optional RR survey. Few of the well-known quantitative RR models are illustrated in ORR context to investigate their degree of protection in privacy. Their performance levels are pointed out in Section 4. The posterior and prior coincide in our proposed quantitative ORR model if the randomized device is either Eichhorn and Hayre (1983) or Chaudhuri 's ***device I***.

Acknowledgements

The authors are grateful to the Editor and referees for their valuable comments and suggestions, which led to considerable improvement in this paper.

References

Anderson, H. (1975a). *Efficiency versus Protection in the RR for Estimating Proportions.* Technical Report 9, University of Lund, Lund, Sweden.

Anderson, H. (1975b). *Efficiency versus Protection in a General RR model.* Technical Report 10, University of Lund, Lund, Sweden.

Anderson, H. (1975c). *Efficiency versus Protection in RR Designs.* Mimeo notes, University of Lund, Lund, Sweden.

Arnab, R. (2004). Optional randomized response techniques for complex designs. *Biometrical Journal*, **46**(**1**), 114-124.

Arnab, R. (2018). Optional randomized response techniques for quantitative characteristics. *Communications in Statistics: Theory and Methods*, **48**(**16**), *4154-4170*.

Chaudhuri, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys.* Boca Raton: CRC Press (ISBN : 978-1-1381-1542-2).

Chaudhuri, A. and Christofides, T.C. (2013). *Indirect Questioning in Sample Surveys.* Berlin, Germany: Springer Verlag (ISBN : 978-3-6423-6275-0).

Chaudhuri, A. and Mukerjee, R. (1985). Optionally randomized responses techniques. *Calcutta Statististical Association Bulletin*, **34**(**3-4**), 225-229.

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

Chaudhuri, A. and Saha, A. (2005). Optional versus compulsory randomized response techniques in complex surveys. *Journal of Statistical Planning and Inference*, **135**(**2**), 516-527.

Chaudhuri, A., Christofides, T.C. and Rao, C.R. (2016). *Handbook of Statistics, Data Gathering, Aanlysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits* (Vol. 34). NL: Elsevier (ISBN : 978-0-4446-3570-9).

Chaudhuri, A., Christofides, T.C. and Saha, A. (2009). Protection of privacy in efficient application of randomized response techniques. *Statistical Methods and Applications*, **18**(**3**), 389-418.

Eichhorn, B. and Hayre, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, **7**(**4**), 307-316.

Eriksson, S.A. (1973). A new model for randomized response. *International Statistical Review*, **41**(**1**), 101-113.

Greenberg, B.G., Kuebler, R.R., Abernathy, J.R. and Horvitz, D.G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of American Statistical Association*, **66**(**334**), 243-250.

Lanke, J. (1975). On the choice of the unrelated question in Simmons' version of randomized response. *Journal of American Statistical Association*, **70**(**349**), 80-83.

Lanke, J. (1976). On the degree of protection in randomized interviews. *International staistical Review*, **44**(**2**), 197-203.

Leysieffer, R.W. and Warner, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of American Statistical Association*, **71**(**355**), 649-656.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, **109**, 325-378.

Pal, S. (2008). Unbiasedly estimating the total of a stigmatizing variable from a complex survey on permitting options for direct or randomized responses. *Statistical Papers*, **49**(**2**), 157-164.

Pollock, K.H. and Bek, Y. (1976). A comparison of three randomized response models for quantitative data. *Journal of American Statistical Association*, **71**(**356**), 884-886.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of American Statistical Association*, **60**(**309**), 63-69.