Statistics and Applications {ISSN 2452-7395 (online)}
Volume 15 Nos. 1&2, 2017 (New Series), pp 19-25

Privacy Protection in Estimating Sensitive Population Proportion by a Modified Unrelated Question Model

Kajal Dihidar and Lagnajita Basu
Indian Statistical Institute, Kolkata, India

Received: December 12, 2016; Revised: January 06, 2017; Accepted: January 30, 2017

Abstract

In this paper we consider the problem of estimating sensitive population proportion by unrelated question model. Motivated by Singh and Grewal (2013), an attempt can be made to develop an inverse unrelated question model. As per the need of successful implementation of the inverse mechanism, a modification of usual unrelated question model is required. In this paper, we propose a modified unrelated question model and derive the unbiased estimator, variance and variance estimators. In addition, we consider here the degree of privacy protection offered to the interviewees. Based on the pioneering work of Leysieffer and Warner (1976), we derive the jeopardy measures for our proposed model. We give a detailed numerical guidance on how to choose the device parameters ensuring the privacy protection within some desired limits as well as maintaining the efficiency in estimation.

Keywords: Unrelated question model, Unbiased estimator.

1 Introduction

Collection of data in surveys on sensitive issues, such as, tax evasion, drug use, illegal abortion, etc. is a very difficult task due to non-cooperation of the respondents, and even if they agree to participate, the truthful answers may not be obtained. To overcome this difficulty, Warner (1965) pioneered the Randomized Response (RR) technique for estimating the proportion of people bearing a stigmatizing attribute, say A in a community, based on a sample of respondents drawn by Simple Random Sampling With Replacement (SRSWR). Since then, many contributors have enriched the randomized response literature by providing alternative models and proving their efficiencies in comparison to the existing techniques, for instance, Horvitz et al. (1967), Greenberg et al. (1969), Kuk (1990), Mangat (1994), Giordano and Perri (2012), Barabesi et. al (2012), Lee et al. (2013) among others. We give some details below.

In Warner's (1965) method, each respondent is provided with a randomization device by which he chooses one of two questions 'Do you belong to A' or 'Do you belong to A^c' with respect to probabilities, say, $p : (1-p)$, where $p \neq 1/2$. The selected respondent is asked to draw randomly one card from the box and is asked to report the 'match' or 'non-match' of his own characteristic with the question written on the card drawn by him. These RR's gathered from a sample of persons provide an unbiased estimator for the sensitive population proportion, say, θ_A. Based on these RRs the variance of this estimator and an unbiased estimator for that variance are also given by Warner (1965).

Later, in order to increase the respondents' participation rate, Horvitz et al. (1967), Greenberg et al. (1969) developed unrelated question method. In this method, along with the questions on sensitive attribute, an additional question on completely unrelated characteristic, say, B, for

Corresponding Author: Kajal Dihidar
Email id:- kajaldihidar@gmail.com

example, 'preference of cricket to football', or 'having the birthday in between January to June', etc is considered. Each respondent selected by a suitable sampling design, say, $p(s)$ is provided with two randomization devices, 'Box1' and 'Box2'. Box1 contains two questions 'Do you belong to A' or 'Do you belong to B' with respect to probabilities, say, $p_1 : (1-p_1)$, and Box2 contains the same questions with respect to probabilities, say, $p_2 : (1-p_2)$ where $p_1 \neq p_2$. The selected respondent is asked to draw randomly one card from Box1 and then independently from Box2 also. He is asked to report the 'match' or 'non-match' of his own characteristic with the questions written on the cards drawn by him. These RR's gathered from a sample of persons provide an unbiased estimator for θ_A.

Recently several researchers have considered generation of the RR's by inverse mechanism (Singh and Grewal (2013), Singh and Sedory (2013), Chaudhuri and Dihidar (2014), Dihidar (2016)). Following them, a problem can be posed on how to estimate the sensitive population proportion by generating RRs following inverse unrelated question model. For this, the randomization device is to be made in such a way as to ensure a positive probability of obtaining the 'match'. The usual unrelated question randomization device does not satisfy this requirement. So, a modification of usual unrelated question randomization device is needed. To fill this need, in this paper, we propose a modified unrelated question model, and based on a sample of respondents drawn by SRSWR, we present the unbiased estimator, variance and variance estimator for sensitive population proportion.

The survey sampling practitioners need to take care of the respondent's privacy to reduce biases due to refusals to respond and intentionally misleading replies while collecting data on sensitive variables. Lanke (1976) studied the issue of respondent's privacy protection and the same issue was studied by Leysiefer and Warner (1976) for dichotomous populations, and by Loynes (1976) for polychotomus populations. Later, Ljungqvist (1993) gave a unified approach to measures of privacy for dichotomous populations, and Nayak and Adeshiyan (2009), Chaudhuri, Christofides and Saha (2009) proposed measures of jeopardy. Recently, Bose (2015) has studied the privacy protection and efficiency for randomized response models for discrete valued sensitive variables. For many other recent rich developments in this direction, we refer to Chaudhuri et al. (2016).

Motivated by these earlier researchers, in this paper, we derive the jeopardy function as suggested in Leysieffer and Warner (1976) for our proposed modified unrelated question model. We find by numerical simulation the optimal design parameters ensuring the privacy protection at some desired level and maintaining maximum efficiency in estimation. We organize our findings of this research work in the following sections.

2 Generating RR by Proposed Modified Unrelated Question Model for Successful Implementation of Inverse Mechanism

Let $U = (1, \ldots, i, \ldots, N)$ denote a finite population of N persons labeled 1 through N. Let

$$\begin{aligned} y_i &= 1 \text{ if } i^{th} \text{ person bears the sensitive characteristic } A \\ &= 0, \text{ otherwise.} \end{aligned}$$

Our objective is to estimate the population proportion $\theta_A = \frac{1}{N}\sum_{i=1}^{N} y_i$ bearing the sensitive characteristic A, using randomized response technique (RRT). To use the unrelated question method, we

define another variable x such that

$$\begin{aligned} x_i &= 1 \text{ if } i^{th} \text{ person belongs to the unrelated characteristic group } B \\ &= 0, \text{ otherwise.} \end{aligned}$$

Let $\theta_B = \frac{1}{N}\sum_{i=1}^{N} x_i$ be the population proportion having the unrelated characteristic B.

Each respondent selected by simple random sampling with replacement is provided with two randomization devices, 'Box1' and 'Box2'. Box1 contains three questions 'Do you belong to A', 'Do you belong to B' and 'Do you belong to B^c' with respect to probabilities, say, $p_1 : p_2 : (1 - p_1 - p_2)$. Box2 contains the same questions with respect to probabilities, say, $p_3 : p_4 : (1 - p_3 - p_4)$ where $\frac{p_1}{1-2p_2} \neq \frac{p_3}{1-2p_4}$. The selected respondent is asked to draw randomly one card from Box1 and then independently from Box2 also. He is asked to report the 'match' or 'non-match' of his own characteristic with the questions written on the cards drawn by him. These RR's gathered from a sample of persons will be used to provide an unbiased estimator for θ_A.

Let λ_1 and λ_2 denote the probability of obtaining 'match' from Box1 and Box2 respectively. Then

$$\lambda_1 = p_1\theta_A + p_2\theta_B + (1 - p_1 - p_2)(1 - \theta_B),$$

and

$$\lambda_2 = p_3\theta_A + p_4\theta_B + (1 - p_3 - p_4)(1 - \theta_B).$$

These two equations yield

$$\theta_A = \frac{a\lambda_1 + b\lambda_2 + c}{d},$$

where

$$a = (p_3 + 2p_4 - 1), \;\; b = (1 - p_1 - 2p_2), \;\; c = (p_1p_4 - p_2p_3 + p_2 - p_4),$$
$$\text{and} \quad d = 2p_4p_1 - 2p_2p_3 + p_3 - p_1.$$

Let n_1 and n_2 denote the number of respondents reporting 'match' from Box1 and Box2 respectively out of n respondents. So,

$$\hat{\lambda_1} = \frac{n_1}{n} \text{ and } \hat{\lambda_2} = \frac{n_2}{n}.$$

On assuming that θ_B as unknown, an unbiased estimator of θ_A can be obtained as

$$\hat{\theta_A} = \frac{a\hat{\lambda_1} + b\hat{\lambda_2} + c}{d},$$

provided $d \neq 0$ which results the condition that $\frac{p_1}{1-2p_2} \neq \frac{p_3}{1-2p_4}$.

Under this condition, the variance of the estimator and unbiased variance estimator are given by

$$V(\hat{\theta_A}) = \frac{a^2\lambda_1(1 - \lambda_1) + b^2\lambda_2(1 - \lambda_2)}{nd^2},$$

$$\hat{V}(\hat{\theta_A}) = \frac{a^2\hat{\lambda_1}(1 - \hat{\lambda_1}) + b^2\hat{\lambda_2}(1 - \hat{\lambda_2})}{(n-1)d^2}.$$

Obviously the variance of the estimator decreases and hence the efficiency increases with the increase in sample size.

3 Privacy Protection of the Proposed Model

We have the possible randomized responses obtained from two boxes as (yes,yes), (yes,no), (no,yes), and (no,no). The conditional probabilities of obtaining those randomized responses obtained from the individual belonging to A or A^c are at the investigator's disposal and are called the design probabilities. For our proposed model, the design probabilities are given by

$$P(yes,yes|A) = \{1-p_2+\theta_B(p_1+2p_2-1)\}\times\{1-p_4+\theta_B(p_3+2p_4-1)\},$$

$$P(yes,yes|A^c) = \{(1-p_1-p_2)+\theta_B(p_1+2p_2-1)\}\times\{(1-p_3-p_4)+\theta_B(p_3+2p_4-1)\},$$

$$P(yes,no|A) = \{(1-p_2)+\theta_B(p_1+2p_2-1)\}\times\{p_4+\theta_B(1-p_3-2p_4)\},$$

$$P(yes,no|A^c) = \{(1-p_1-p_2)+\theta_B(p_1+2p_2-1)\}\times\{(p_3+p_4)+\theta_B(1-p_3-2p_4)\},$$

$$P(no,yes|A) = \{p_2+\theta_B(1-p_1-2p_2)\}\times\{(1-p_4)+\theta_B(p_3+2p_4-1)\},$$

$$P(no,yes|A^c) = \{(p_1+p_2)+\theta_B(1-p_1-2p_2))\}\times\{(1-p_3-p_4)+\theta_B(p_3+2p_4-1)\},$$

$$P(no,no|A) = \{p_2+\theta_B(1-p_1-2p_2)\}\times\{p_4+\theta_B(1-p_3-2p_4)\},$$

and

$$P(no,no|A^c) = \{(p_1+p_2)+\theta_B(1-p_1-2p_2)\}\times\{(p_3+p_4)+\theta_B(1-p_3-2p_4)\}.$$

Following Leysieffer and Warner (1976), we consider the measure of jeopardy carried by a randomized response 'yes' or 'no' about A and A^c , respectively. For our proposed model, we define these measures as follows:

$$g(yes|A) = \frac{P(yes,yes|A)+P(yes,no|A)+P(no,yes|A)}{P(yes,yes|A^c)+P(yes,no|A^c)+P(no,yes|A^c)} = \frac{1-P(no,no|A)}{1-P(no,no|A^c)},$$

and

$$g(yes|A^c) = \frac{1}{g(yes|A)}.$$

Similarly,

$$g(no|A) = \frac{P(no,no|A)+P(no,yes|A)+P(yes,no|A)}{P(no,no|A^c)+P(no,yes|A^c)+P(yes,no|A^c)}$$

$$= \frac{1 - P(yes, yes|A)}{1 - P(yes, yes|A^c)},$$

and

$$g(no|A^c) = \frac{1}{g(no|A)}.$$

So, if $g(yes|A) > 1$, which reduces to the inequality that if $P(no, no|A) < P(no, no|A^c)$, then a 'yes' answer increases the odds of A and is jeopardizing with respect to A, while a 'no' answer increases the odds of A^c and is jeopardizing with respect to A^c, if $g(no|A^c) > 1$ or in other words if $P(yes, yes|A) > P(yes, yes|A^c)$.

4 Privacy Protection and Efficiency in Estimation

We now simultaneously consider the two issues of efficiency in estimation and protection of respondent privacy in randomized response surveys using the proposed modified unrelated question model. From the preceding section, we note that a very high value of $g(yes|A)$ indicates a very high risk of "yes" answer jeopardizing with respect to A, and similarly, very high value of $g(no|A^c)$ indicates a very high risk of "no" answer jeopardizing with respect to A^c. So, we would like to choose the parameters of the randomization device, in such a way so as to achieve high efficiency of estimation while ensuring a desired level of privacy protection, i.e. controlling jeopardy measures within our chosen limits. Below we give a detailed guideline how to choose the device parameters for our proposed modified unrelated question model, allowing the jeopardy measures up to a maximum of 2, by considering the desired class limits, say, 1 to 1.2, 1.2 to 1.4,...., 1.8 to 2.0, and at the same time maintaining the maximum efficiency in estimation. In this regard, for numerical illustration we consider the sample size 100, and for each combination of $\theta_A = 0.1, 0.2, ..., 0.9$, and $\theta_B = 0.1, 0.2, ..., 0.9$, we consider the ranges of $g_1 = g(yes|A)$ and $g_2 = g(no|A^c)$ within the limits (1, 1.2], (1.2, 1.4], (1.4, 1.6], (1.6, 1.8], (1.8, 2.0].

In Table 1 which an be viewed as a pdf file by linking here (url of Table 1:https://drive.google.com/open?id=0Bwfbv7zaZGB6cGhiX19ralc4Q3c), we present the values of the device parmeters p_1, p_2, p_3, and p_4 resulting in the minimum and maximum variances in estimation. However, in practice the values of θ_A and θ_B will be unknown. So to take advantage of this model, following Mangat (1994), Lee et al. (2013) among others, the interviewer may use these two proportion values based on some prior guess obtained from an earlier study or pilot survey. It is now clear that, following the results in Table 1 the use of randomized response technique with the proposed modified unrelated question model will be able to choose the device parameters keeping the jeopardy limits within his own choices and thus will be able to keep the level of privacy protection within his own desired limits. Furthermore, he will be aware about the expected efficiency in estimation. Hence the justification of this research.

5 Concluding Remarks

In this paper we propose a modified unrelated question randomization device for collecting randomized response data on binary sensitive variable from a sample of individuals, and present the related estimation procedures. Our proposed device is suitable for randomized response data collection by inverse mechanism. At the same time, we concentrate on quantifying the extent of

privacy protection available to the respondents when they give responses under this scheme. For this, following Leysieffer and Warner (1976) we derive the jeopardy measures for our proposed model. We give a detailed guideline by a numerical presentation that for some maximum allowable limits of jeopardy measures, how parameters of the device can be chosen so that we can guarantee the level of protection to lie within some fixed limits and also obtain efficient estimates of the sensitive population proportion. Thus, it is expected that the interviewers will be able to convince the sampled respondents and reassure them as to their privacy protection. Thus we may expect respondents to readily participate in the survey, resulting in efficient estimators.

Acknowledgements

The authors are grateful to an anonymous referee for his/her constructive and insightful comments which led to improvements over the earlier manuscript.

References

Anderson, H. (1977). Efficiency versus protection in a general randomized response model. *Scandinavian Journal of Statistics.* **4**, 11-19.

Barabesi, L., Franceschi, S. and Marcheselli, M. (2012). A randomized response procedure for multiple-sensitive questions. *Statistical Papers.* **53**, 703-718.

Bose, M. (2015). Respondent privacy and estimation efficiency in randomized response surveys for discrete-valued sensitive variables. *Statistical Papers*. **4**, 1055-1069.

Chaudhuri, A., Christofides, T. C. and Rao, C.R. (Eds). (2016). *Handbook of Statistics **Vol. 34**, Data gathering, analysis and protection of privacy through randomized response techniques.* Amsterdam : Elsevier.

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

Chaudhuri, A. and Dihidar, K. (2014). Generating randomized response by inverse mechanism. *Model Assisted Statistics and Applications.* **9**, 343-351.

Dihidar, K. (2016). Estimating Sensitive Population Proportion by Generating Randomized Response Following Direct and Inverse Hypergeometric Distribution. In Chaudhuri, A., Christofides, T. C. and Rao, C.R. (Eds). *Handbook of Statistics **Vol. 34, pp - 427-441**, Data gathering, analysis and protection of privacy through randomized response techniques.* Amsterdam : Elsevier. .

Giordano, S. and Perri, P. F. (2012). Efficiency comparison of unrelated question models based on same privacy protection degree. *Statistical Papers.* **53**, 987-999.

Greenberg, B.G., Abul-Ela, Abdel-Latif, A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question RR model : theoretical framework. *Journal of American Statistical Association.* **64**, 520-539.

Horvitz, D.G., Shah, B.V. and Simmons, W.R. (1967). The unrelated question randomized response model. *Social Statistics Section, Proceedings of the American Statistical Association.* 65-72.

Kuk, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika.* **77**, 436-438.

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

Lee, C-S., Sedory, S.A. and Singh, S. (2013). Simulated minimum sample size requirements in various randomized response models. *Communications in Statistics : Simulation and Computation.* **42**, 771-789.

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

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

Loynes, R.M. (1976). Asymptotically optimal randomized response procedures. *Journal of American Statistical Association.* **71**, 924-928.

Mangat, N.S. (1994). An improved randomized response strategy. *Journal of Royal Statistical Society, Series B.* **56**, 93-95.

Nayak, T. K. and Adeshiyan, S. A. (2009). A unified framework for analysis and comparison of randomized response surveys of binary characteristics. *Journal of Statistical Planning and Inference.* **139**, 2757-2766.

Singh, S. and Grewal, I.S. (2013). Geometric distribution as a randomization device: Implemented to the Kuks model. *International Journal of Contemporary Mathematical Sciences.* **8(5)**, 243-248.

Singh, S. and Sedory, S.A. (2013). A new randomized response device for sensitive characteristics: An application of negative hypergeometric distribution. *Metron.* **71**, 3-8.

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