Measures of privacy in randomized response surveys for quantitative stigmatizing variables

Mausumi Bose

Indian Statistical Institute, Kolkata 700108, India
email: mausumi@isical.ac.in

Abstract

In many socio-economic surveys, the variable of interest is sensitive or stigmatizing. Examples include tax evasion, criminal conviction, alcohol expenses, induced abortion, drug abuse, etc. In such situations, the technique of randomized response is very useful as it does not require the respondent to reveal his/her true value. The issue of privacy protection is important in this context and a few researchers have studied this problem for surveys on dichotomous populations, where the objective is to estimate the proportion of persons bearing the sensitive trait. There is a considerable literature on various randomized response techniques. However, not much is known as yet about the extent of privacy protection when the variable under study is quantitative in nature. In this article we study the issue of privacy protection when the randomized response technique is used for a quantitative variable which could be either discrete or continuous. We propose a measure of privacy for each case. Taking cognizance of a conflict between protection of privacy and enhancing estimation efficiency, we discuss how, given a stipulated level of our privacy measure, the parameters of the randomization device can be determined so as to maximize the efficiency of estimation. Numerical examples are provided to illustrate our results.

Keywords: Continuous sensitive variable, discrete sensitive variable, revealing distribution, distance measure.

1 Introduction

In some surveys we need to gather information on variables which are sensitive or stigmatizing in nature, for instance, we may want to estimate the proportion of persons who have been convicted of a certain crime or have evaded tax, or estimate the extent of undisclosed income, etc. In such surveys, direct questions are not useful as the respondent will either refuse to answer embarrassing questions or will give false answers. The randomized response technique is useful in such cases since in this method, a respondent uses a randomization device to generate a randomized response without revealing his/her true response; thereby increasing participation in the survey. From these randomized responses, the parameter under study can be estimated.

Warner (1965) introduced this technique for estimating the proportion of persons bearing a sensitive attribute in a population, based on a simple random sample of individuals chosen from the population with replacement. Subsequently, Kuk (1990), Ljungqvist (1993), Chua and Tsui (2000), Van den Hout and Van der Heijden (2002), Christofides (2005) and many others

have contributed to this area and modified Warner's (1965) technique in several directions. For details on all results available on this technique we refer to the books by Chaudhuri and Mukerjee (1988) and Chaudhuri (2011).

It is clear that in a randomized response survey, the privacy of respondents is protected since they report their response after making use of a randomization device. There has been some work in studying the degree of protection available to respondents vis-a-vis the efficiency of estimation from the randomized responses. Lanke (1976) and Leysieffer and Warner (1976) introduced this study of privacy protection for dichotomous populations, i.e., populations which consist of two types of individuals, those that bear the stigmatizing characteristic and those that do not, and the interest is to estimate the proportion of persons with the stigmatizing characteristic. Loynes (1976) extended the results to polychotomous populations. Nayak and Adeshiyan (2009) gave a unified framework for studying this problem for dichotomous populations. Chaudhuri et al. extended the work of Nayak and Adeshiyan (2009) to sampling with varying probability. It may be noted that all these studies are applicable to only qualitative sensitive variables.

Not much work seems to be have been done for the situation where the study variable is quantitative in nature, even though this is a common situation, e.g. in studies to estimate the number of convictions in the past year, the number of induced abortions, or the amount of undisclosed income, etc. The only work known to us in this area is by Anderson (1977) who studied this problem for continuous stigmatizing variables. No results seem to be available for discrete valued variables.

In this article we study the privacy protection aspect when the underlying variable under study is quantitative. We study both discrete and continuous variables, and propose measures for protecting the privacy under each case separately. It may be noted that our study for the discrete sensitive variable also covers the situation where randomized response technique is popularly used, namely, estimation of proportions of persons bearing a sensitive attribute in a population. We show that, for a stipulated level of privacy protection, it is possible to choose a randomization device so as to guarantee this protection and also ensure efficient estimation. We use simple random sampling without replacement to select the sample.

In Section 2 we give some preliminaries. In Sections 3 and 4 we consider the discrete case and continuous case, respectively.

2 Discrete sensitive character

2.1 Preliminaries

Let X denote the sensitive variable of interest, which is discrete in nature. The objective of the survey is to estimate the mean value of X for the population under study. For this, we suppose that the population consists of N individuals labeled $1, \ldots, N$ and a sample of n individuals is drawn from this population by SRSWOR.

We assume that X takes a finite number of possible values $x_1, \ldots, x_m$, not all these values necessarily occurring in the population. Without loss of generality, we assume these values are known. Suppose each sampled individual is given a randomization device and and asked to report a random-

ized response after using this device. Let R denote the randomized response variable and ideally, the device should be such that the possible randomized responses should match with the possible values of the true response. i.e., R takes the values $x_1, \ldots, x_m$.

For any respondent, let p_{ij} denote the probability that a respondent for whom the true value of X is x_i, reports a randomized response as x_j, $\sum_{i=1}^{m} p_{ij} = 1$ for all j, $1 \leq i, j \leq m$.

Let π_i be the unknown proportion of persons in the population for whom the value of X is equal to x_i, $1 \leq i \leq m$, with $\pi_i \geq 0$, $\sum_{i=1}^{m} \pi_i = 1$. Let $\mathbf{x} = (x_1, \ldots, x_m)'$, and $\pi = (\pi_1, \ldots, \pi_m)'$ be $m \times 1$ vectors.

Then expressions for the population mean and variance of X can be obtained in terms of these probabilities and proportions.

Suppose a SRSWOR of size n is drawn from the population and the randomized response for each sampled individual is recorded. Our objective is to estimate μ_X.

From the randomized responses, for the randomization device under use, we may obtain expressions for estimate of μ and its variance.

Remark 2.1 The case of qualitative variables is also covered by our discussion on discrete variables. For example, suppose the population is dichotomous, the sensitive variable under study is a qualitative attribute and we want to estimate the proportion θ of persons who bear this attribute. Clearly, if in the above study we put $m = 2$ with $x_1 = 0, x_2 = 1$, then μ is equal to θ and so estimating the proportion is same as estimating μ as above. Again, if the variable has a number of classes and we want to estimate the proportions of persons in these classes, this estimate also follows from the above study. □

2.2 Protection of privacy

The issue of privacy protection in randomized response surveys is important since higher the level of protection, the better will be the level of participation of individuals in the study. To study the amount of privacy protection available under this method one needs to study how much information can be extracted about the sensitive variable from the available randomized responses.

We develop a measure for measuring the level of privacy protection available to respondents under a randomization scheme. We show how the device may be used to collect responses such that the efficiency of estimation can be kept at a high level subject to the condition that the level of privacy if above a certain required level.

While designing a randomization device, one seeks to keep the efficiency of estimation as high as possible and at the same time safeguard the privacy of the respondents as much as possible. Suppose a lower bound for the level of protection required is stipulated and then a scheme is needed such that the parameter of interest can be efficiently estimated while ensuring this level of protection. We give a method such that the device parameter can be estimated efficiently and the privacy protection is also kept above a certain level.

3 Continuous sensitive variable

Consider a population with N individuals and let S denote the sensitive variable of interest, which is continuous in nature. The objective of the survey is to estimate the population mean of S. For this, we suppose that a sample of n individuals is drawn from the population by simple random sampling without replacement (SRSWOR). Suppose each sampled individual is given a randomization device and and asked to report a randomized response after using this device.

In RR techniques with a continuous variable of interest, a scrambling variable with a known distribution is commonly used to generate the responses. Let A be such a scrambling variable and suppose on using a randomization device, we get a randomized response R as

$$R = AS. \tag{1}$$

Several devices can be thought of which will allow the respondent to respond as in (1). As in Section 2 we would prefer the device to be such that the range of the randomized response variable R matches that of Y.

We now develop a method for estimation of the parameter of interest and also develop a measure for privacy protection. We show how given a certain required level of protection, this level can be achieved and at the same time an efficient estimation of the population mean can be done.

References

Anderson, H. (1977) Efficiency versus protection in a general randomized response model. Scand. J. Statist. 4, 11-19.

Chaudhuri, A. (2011) *Randomized response and indirect questioning techniques in surveys.* CRC Press, Boca Raton, FL.

Chaudhuri, A. and Mukerjee, R. (1987) Randomized response techniques: a review. Statist. Neerlandica 41 1, 27-44.

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

Christofides, T.C., (2005) Randomized response in stratified sampling. J. Statist. Plann. Inference 128, 303-310.

Chua, T.C. and Tsui, A.K. (2000) Procuring honest responses indirectly. J. Statist. Plann. Inference 90, 107-116.

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

Lanke, J. (1976) On the degree of protection in randomized interviews. Int. Stat. Rev. 44, 197–203.

Leysieffer, R.W. and Warner, S.L. (1976) Respondent jeopardy and optimal designs in randomized response models. J. Amer. Statist. Assoc. 71, 649–656.

Ljungqvist, L. (1993) A unified approach to measures of privacy protection in randomized response models: a utilitarian perspective. J. Amer. Statist. Assoc. 88, 97–103.

Loynes, R.M. (1976) Asymptotically Optimal Randomized Response Procedures. J. Amer. Statist. Assoc. 71, 924–928.

Nayak, T. K. and Adeshiyan, S. A. (2009) A unified framework for analysis and comparison of randomized response surveys of binary characteristics. J. Statist. Plann. Inf. 139, 2757–2766.

Van den Hout, A. and Van der Heijden, P.G.M. (2002) Randomized response, statistical disclosure control and misclassification: a review. Internat. Statist. Rev.70, 269-288.

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