Stat Papers (2015) 56:1055–1069
DOI 10.1007/s00362-014-0624-4

REGULAR ARTICLE

Respondent privacy and estimation efficiency in randomized response surveys for discrete-valued sensitive variables

Mausumi Bose

Received: 18 September 2013 / Revised: 24 December 2013 / Published online: 19 August 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract In some socio-economic surveys, data are collected on sensitive issues such as tax evasion, criminal conviction, drug use, etc. In such surveys, direct questioning of respondents is not of much use and the randomized response technique is used instead. A few researchers have studied the issue of privacy protection for surveys where the objective is to estimate the proportion of persons bearing the sensitive trait. Not much is known about respondent protection when the variable under study is a discrete quantitative variable and the objective is to estimate the population mean. In this article we study this issue. We propose a scheme for this issue and a measure of privacy. We show that given a stipulated level of this privacy measure, we can determine the parameter of the randomization device so as to maximize the efficiency of estimation, while guaranteeing the desired level of privacy protection.

Keywords Jeopardy measure · Numerical stigmatizing variable · Revealing probability · Socio-economic sample surveys

MSC Classification 62D05

1 Introduction

The randomized response technique is a useful method for collecting data on variables which are considered sensitive, incriminating or stigmatizing for the respondents. Examples of such situations are common in socio-economic surveys, for instance, we may need to collect data on tax evasion, alcohol addiction, illegal drug use, criminal behaviour or past criminal convictions. In such surveys, direct questions are not useful

M. Bose (✉)
Indian Statistical Institute, Kolkata 700108, India
e-mail: mausumi.bose@gmail.com

as the respondents will either refuse to answer embarrassing questions or, even if they do, may give false answers. In a randomized response model, the respondents use a randomization device to generate a randomized response and the parameter under study can be estimated from these responses. So, the respondent is not required to disclose his true response and it is expected that this will lead to better participation in the survey on sensitive issues.

Warner (1965) introduced the randomized response technique for estimating the proportion of persons in a dichotomous population bearing a sensitive qualitative character, such as alcoholism or drug addition. Let A denote such a sensitive character and A^c its complement. Suppose in the population, the proportion of individuals bearing the character A is π_A. Then the proportion bearing A^c is $\pi_{A^c} = 1 - \pi_A$. The objective of the survey is to estimate the unknown value π_A. In Warner's randomized response model, a box with two types of cards labeled A and A^c (in proportion $p : 1 - p$, $p \neq 1/2$) is used as the randomization device. Each respondent is asked to draw a card at random from the box and then respond simply 'yes' or 'no' according as whether or not he bears the character on the label of the card he draws, without disclosing this label. Thus the actual character of the respondent is not disclosed. If this randomization procedure is adopted, on the basis of a simple random sample with replacement of size n, the proportion π_A may be unbiasedly estimated by $\{w(1 - p)\}/(2p - 1)$, where w is the sample proportion of 'yes' responses. Thus π_A may be unbiasedly estimated without having recourse to direct questioning.

Since then, many researchers have extensively contributed to this area, some have proposed alternative models for estimating proportions in dichotomous populations, some have extended this to proportions in polychotomous models and others have studied the case of quantitative sensitive variables, e.g., Kuk (1990); Ljungqvist (1993); Mangat (1994); Chua and Tsui (2000); Van den Hout and Van der Heijden (2002); Christofides (2005); Kim (2007); Arnab and Dorffner (2007); Pal (2008); Diana and Perri (2009);(Chaudhuri et al. (2011a); Chaudhuri et al. (2011b)); Barabesi et al. (2012) and many others. For details on the results available on this technique we refer to the review paper by Chaudhuri and Mukerjee (1987) and books by Chaudhuri and Mukerjee (1988) and Chaudhuri (2011).

Lanke (1976) and Leysieffer and Warner (1976) initiated the study of efficiency versus privacy protection in randomized response surveys where the population is divided into two complementary sensitive groups, A and A^c, and the objective is to estimate the proportions of persons belonging to these two groups. They suggested measures of jeopardy based on the 'revealing probabilities', i.e., the posterior probabilities of a respondent belonging to groups A and A^c given his randomized response. Since then, this dichotomous case has been widely studied. Loynes (1976) extended the jeopardy measure of Leysieffer and Warner (1976) to polychotomous populations. Ljungqvist (1993) gave a unified and utilitarian approach to measures of privacy for the dichotomous case. For estimating the proportion π_A, Nayak and Adeshiyan (2009) proposed a measure of jeopardy for surveys from dichotomous populations and developed an approach for comparing the available randomization procedures. Recently, for estimating π_A, Giordano and Perri (2012) compared some randomization models from the point of view of efficiency and privacy protection.

All the references given above are for sensitive variables which are categorial or qualitative in nature and the objective is to estimate π_A. However, in randomized response surveys it is quite common to have situations where the study variable X is quantitative, e.g. in studies on the number of criminal convictions of a person, the number of induced abortions, the amount of time spent in a correction centre, the amount of undisclosed income, etc. Anderson (1977) studied the case of continuous sensitive variables and considered the amount of information provided by the randomized responses. For ensuring more privacy he recommended that the expectation of the conditional variance of X given the randomized response be made as large as possible. Diana and Perri (2011) studied quantitative sensitive data and for estimating the mean, they used auxiliary information at the estimation stage and compared different models from the efficiency and privacy protection aspects. However, notwithstanding the rich literature on the randomized response technique, not much work seems to have been done in studying the respondent privacy aspect for discrete-valued sensitive variables, even though surveys are often undertaken on such variables.

To fill this gap, in this article we focus on studying the issue of privacy protection when the underlying sensitive variable under study is quantitative and discrete. For instance, the variable of interest may be the number of convictions for criminal offences, number of times one has used illegal drugs or the number of induced abortions, etc. We propose the use of a randomization device and give the associated estimation method for such studies. Then, we consider two separate cases, one where all values of X are sensitive and another where not all values of X are sensitive. For each of these cases, we propose a measure for protecting the privacy of the respondents. We finally show how one can choose the randomization device parameter in each case, so as to guarantee a certain pre-specified level of respondent protection and then maximize the efficiency of estimating the parameter of interest under this constraint. Our study also covers qualitative sensitive variables, i.e., cases where the population is dichotomous or polychotomous, and allows us to estimate the proportions of individuals belonging to each category.

In Sect. 2 we give some preliminaries. In Sect. 3 and 4 we consider the issues of estimation and privacy protection, respectively. In Sect. 5 we obtain the randomization device parameter which allows efficient estimation while assuring the required level of respondent protection. We also present some illustrative numerical examples. In Sect. 6 we show how our study covers the case of polychotomous variables. Finally we conclude with some remarks.

2 Preliminaries

Consider a population of individuals and let X denote the sensitive variable of interest. We assume that X takes a finite number of values $x_1, \ldots, x_m$ and without loss of generality, we may suppose these m values to be known. For $1 \leq i \leq m$, let π_i be the unknown population proportion of individuals for whom X equals x_i, i.e.,

$$\text{Prob}(X = x_i) = \pi_i, \quad 1 \leq i \leq m, \quad \text{where } \pi_i \geq 0, \quad \sum_{i=1}^{m} \pi_i = 1. \tag{1}$$

The objective of the survey is to estimate the population mean of X. For this, we suppose as usual (cf. Warner (1965), Nayak and Adeshiyan (2009) and others), that a sample of n individuals is drawn from the population by simple random sampling with replacement. As for the randomization device, since we are interested in the numerical values of X, we propose the use of a device as described below.

Consider a box containing cards of $(m+1)$ types, the ith type of card being marked 'Report x_i as your response', $1 \le i \le m$, while the $(m+1)$th type of card is marked: 'Report your true value of X as your response.' The box has a large number of cards, say M, there being Mp cards of type $(m+1)$ and $M\frac{1-p}{m}$ cards of each of the types i, $1 \le i \le m, 0 < p < 1$. A sampled respondent is asked to draw a card at random from the box and then give a truthful response according to the card drawn by him, without disclosing the label on the card to the investigator. Thus the true value of X for the respondent is not known. The n responses so received are the data from this survey.

Let R denote the randomized response variable. Clearly, with this device, the ranges of R and X match. The efficiency in estimation and respondent protection will depend on the choice of the value of p, which we call the device parameter. The above device is such that with probability p, a respondent will report his true value, while with probability $\frac{1-p}{m}$, he will report any one of the possible values $x_1, \ldots, x_m$ chosen at random, i.e.,

$$\text{Prob}(R = x_i | X = x_j) = \frac{1-p}{m}, \quad 1 \le i \ne j \le m, \tag{2}$$

$$\text{Prob}(R = x_j | X = x_j) = p + \frac{1-p}{m}, \quad 1 \le j \le m. \tag{3}$$

3 Estimation of population mean

The population mean and variance of X are given by

$$\mu_X = \sum_{i=1}^{m} x_i \pi_i \quad \text{and} \quad \sigma_X^2 = \sum_{i=1}^{m} (x_i - \mu_X)^2 \pi_i,$$

respectively. Our objective is to estimate μ_X from the n randomized responses collected as described in Sect. 2. Let w_i be the sample proportion of randomized responses which equal x_i, $1 \le i \le m$. Hence, from (1)-(3),

$$\text{E}(w_i) = \text{Prob}(R = x_i) = p\pi_i + \frac{1-p}{m} = \lambda_i, \quad \text{say.} \tag{4}$$

So, an unbiased estimator of π_i will be given by $\hat{\pi}_i = \frac{1}{p}(w_i - \frac{1-p}{m})$, leading to an unbiased estimator of μ_X as

$$\hat{\mu}_X = \sum_{i=1}^{m} x_i \hat{\pi}_i = \frac{1}{p} \sum_{i=1}^{m} x_i w_i - \frac{1-p}{mp} \sum_{i=1}^{m} x_i.$$

Let us write $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} x_i$. Then, on simplification using (4), the variance of $\hat{\mu}_X$ for a given value of p is given by

$$\begin{aligned}\mathrm{Var}_p(\hat{\mu}_X) &= \frac{1}{p^2}\mathrm{Var}\Big(\sum_{i=1}^{m} x_i w_i\Big) = \frac{1}{np^2}\left\{\sum_{i=1}^{m} x_i^2\lambda_i(1-\lambda_i) - \sum_{i=1}^{m}\sum_{j(\neq i)=1}^{m} x_i x_j \lambda_i \lambda_j\right\} \\ &= \frac{1}{np^2}\left\{p\sum_{i=1}^{m} x_i^2\pi_i + \frac{1-p}{m}\sum_{i=1}^{m} x_i^2 - \big(p\mu_X + (1-p)\bar{X}\big)^2\right\} \\ &= \frac{1}{np^2}\left\{p\sigma_X^2 + (1-p)\frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{X})^2 + p(1-p)(\mu_X - \bar{X})^2\right\}. \quad (5)\end{aligned}$$

Our aim is to estimate μ_X keeping $\mathrm{Var}_p(\hat{\mu}_X)$ as small as possible. It is clear from the expression on the right side of (5) that $\mathrm{Var}_p(\hat{\mu}_X)$ is decreasing in p, irrespective of the values of $\pi_1, \ldots, \pi_m$. So, this variance may be decreased, or equivalently, the efficiency of estimation may be increased by increasing p, whatever may be the proportions of the x_i values in the population.

4 Privacy protection

In this section we consider the degree of privacy protection available to respondents in a randomized response survey where the randomization device is as described in Sect. 2. In the literature, while studying the respondent privacy aspect for dichotomous populations, Leysieffer and Warner (1976) studied the case where both A and A^c are sensitive categories while Lanke (1975) also considered the case where only A is sensitive and there is no jeopardy in a 'no' answer to the sensitive question. For polychotomous populations, Loynes (1976) studied two cases, one where all categories are stigmatizing and another where one of the categories is not stigmatizing.

In line with these studies for qualitative stigmatizing variables, we too consider the privacy issue for discrete-valued variables for two situations, one where all the m values of X are stigmatizing and another where not all values of X are stigmatizing. For instance, we may want to study the number of times a person has voted for a certain political party in the last 5 elections. Here possible values of X are $X = 0, \ldots, 5$, all of which may be sensitive. On the other hand, if we are studying the number of times one has under-reported his income for income tax, the value $X = 0$ is not stigmatizing, but any value of X larger than zero may well be sensitive. Again, if one is studying the number of induced abortions, then values $X = 0$ or $X = 1$ may not be considered stigmatizing but higher values of X may be deemed to be so. Many such examples may be cited to show that both these situations arise commonly in practice. We will show that we require separate privacy protection measures in these cases.

For a randomly chosen respondent from the population, the 'true' probability that the value of X for this respondent equals x_i is given by $\mathrm{Prob}(X = x_i)$. On the other hand, when this respondent gives a randomized response, say x_j, then the probability that the value of X for this respondent equals x_i is now given by the conditional

probability Prob$(X = x_i|R = x_j)$, or the 'revealing' probability. A respondent will be assured that his privacy is protected if he can be convinced that given his response, the probability of his having a particular value of X does not change much, i.e., he needs to be assured that the difference between his true probability and his revealing probability is as small as possible, for all possible true values and all responses. With this is mind, in the next subsections we develop measures of privacy protection separately for the different cases. Further, in a given situation, for a certain target level of privacy protection by the relevant measure, we obtain a range of values of p for the randomization device which can achieve this.

4.1 All values of X are stigmatizing

Suppose all the values $x_1, \ldots, x_m$ are stigmatizing. In this case, a respondent would feel comfortable in participating in the survey if the perception of his having a value $X = x_i$ is not much altered after knowing his randomized response, for all $1 \leq i \leq m$. This would require that his true and revealing probabilities be sufficiently close. Starting from this basic premise we define

$$\alpha_{ij} = |Prob(X = x_i|R = x_j) - Prob(X = x_i)| \tag{6}$$

and since each respondent would want α_{ij} to be as small as possible for all $1 \leq i, j \leq m$, as a measure of privacy protection we propose the following measure:

$$\alpha = \max_{1 \leq i, j \leq m} \alpha_{ij}. \tag{7}$$

A randomization device with a privacy protection value $\alpha = \alpha_0$ would guarantee that the discrepancies between the true and revealing probabilities will be at most α_0 for all respondents, irrespective of their true values. Thus a device which results in a lower value of α gives a higher level of privacy protection than one with a higher value of α.

Suppose the scientist planning a certain survey would like to keep the privacy protection available to respondents above a certain threshold, i.e., would like to achieve $\alpha \leq L$, where L is a pre-assigned quantity, $0 < L < 1$. Moreover, this bound on α should hold irrespective of the unknown values of $\pi_1, \ldots, \pi_m$. The following theorem shows how the device parameter can be chosen to achieve this.

Theorem 1 *For α as in (7) and a preassigned L, where $0 < L < 1$, $\alpha \leq L$ will hold, irrespective of the values of $\pi_1, \ldots, \pi_m$, if and only if $p \leq p_0$, where*

$$p_0 = \frac{1}{1 + \frac{m}{L}(\frac{1-L}{2})^2}. \tag{8}$$

Proof From (1)-(3), using Bayes' Theorem it follows that for $1 \leq i, j, \leq m$,

$$Prob(X = x_i|R = x_j) = \frac{(p\delta_{ij} + \frac{1-p}{m})\pi_i}{\sum_{u=1}^{m}(p\delta_{ju} + \frac{1-p}{m})\pi_u} = \frac{(p\delta_{ij} + \frac{1-p}{m})\pi_i}{p\pi_j + \frac{1-p}{m}}, \tag{9}$$

where δ_{ij} is Kronecker Delta. Hence from (6) it follows that $\alpha_{ij} = \frac{p\pi_i|\pi_j - \delta_{ij}|}{p\pi_j + \frac{1-p}{m}}$ and for any $i \neq j$,

$$\alpha_{ij} = \frac{p\pi_i\pi_j}{p\pi_j + \frac{1-p}{m}} \leq \frac{p(1-\pi_j)\pi_j}{p\pi_j + \frac{1-p}{m}} = \alpha_{jj}, \tag{10}$$

as $\pi_i + \pi_j \leq 1$ for all i, j. Thus $\alpha = \underset{1\leq j\leq m}{max}\ \alpha_{jj} = \underset{1\leq j\leq m}{max} \frac{\pi_j(1-\pi_j)}{\pi_j + \frac{1-p}{mp}}$. Hence, $\alpha \leq L$ if and only if

$$\pi_j(1-\pi_j) - L\pi_j \leq \frac{L(1-p)}{mp} \quad for\ \ all\ \ 1 \leq j \leq m. \tag{11}$$

First suppose $p \leq p_0$. Then for $1 \leq j \leq m$,

$$\begin{aligned}
\pi_j(1-\pi_j) - L\pi_j &= \left(\frac{1-L}{2}\right)^2 - \left(\frac{1-L}{2} - \pi_j\right)^2 \\
&\leq \left(\frac{1-L}{2}\right)^2 \\
&= \frac{L(1-p_0)}{mp_0}, \quad using\ the\ expression\ of\ \ p_0\ in\ \ (8) \\
&\leq \frac{L(1-p)}{mp}, \quad since\ \ p \leq p_0.
\end{aligned}$$

Thus the inequalities in (11) hold, or equivalently $\alpha \leq L$, irrespective of the values of $\pi_1, \ldots, \pi_m$.

To prove the converse, suppose $\alpha \leq L$, or equivalently, the inequalities in (11) hold, irrespective of the values of $\pi_1, \ldots, \pi_m$. Then, for $\pi_1 = \frac{1-L}{2}, \pi_2 = \frac{1+L}{2}, \pi_3 = \ldots = \pi_m = 0$, in particular, these inequalities will also hold. So, for this choice of π_j values in (11) with $j = 1$, we have

$$\begin{gathered}
\left(\frac{1-L}{2}\right)\left(\frac{1+L}{2}\right) - L\left(\frac{1-L}{2}\right) \leq \frac{L(1-p)}{mp} \\
\text{i.}e.,\ \ \left(\frac{1-L}{2}\right)^2 \leq \frac{L(1-p)}{mp}, \\
\text{i.}e.,\ \ \frac{m}{L}\left(\frac{1-L}{2}\right)^2 \leq \frac{(1-p)}{p} = \frac{1}{p} - 1 \\
\text{i.}e.,\ \ \frac{1}{p_0} \leq \frac{1}{p}, \quad using\ the\ expression\ of\ \ p_0\ in\ \ (8).
\end{gathered} \tag{12}$$

Hence $p \leq p_0$. Hence theorem. □

Remark 1 It is clear from (8) that in order to maintain the same level of protection, the value of p_0 monotonically decreases with the number of possible values of X. Again, for a given number of possible values of X, p_0 monotonically increases with

L. We may reiterate that these values of p do not depend on how the values of X are distributed in the population.

4.2 Not all values of X are stigmatizing

In many surveys it may so happen that not all values of X are sensitive or stigmatizing. For instance, in a survey for estimating the average number of criminal convictions of persons in a certain population, the value $X = 0$ is not stigmatizing but any value of $X \geq 1$ could well be stigmatizing. Similarly, for a survey for estimating the average of the number (X) of induced abortions, the values $X = 0$ or $X = 1$ might not be considered as stigmatizing values while other larger values might be considered stigmatizing by the respondents.

To study the respondents' privacy protection for such surveys, we first present the simpler case where only one of the values of X, say x_1, is not stigmatizing, while values $x_2, \ldots, x_m$ are considered stigmatizing. We develop the protection measure for this case in detail. Later we remark that the results obtained for this case may be easily extended to the case where X has more than one non-stigmatizing value.

As before, the data collection and estimation proceeds as in Sect. 2 and 3. To study the respondent protection we note that since the value x_1 is non-stigmatizing, respondents will feel comfortable with a randomization device for which the 'revealing' probability of their having a true value x_1 will be large. So, we propose the following measure of privacy:

$$\beta = \min_{1 \leq j \leq m} P(X = x_1 | R = x_j) = \min_{1 \leq j \leq m} \frac{(p\delta_{1j} + \frac{1-p}{m})\pi_1}{p\pi_j + \frac{1-p}{m}}, \tag{13}$$

on simplification using (9). A device with a privacy protection value β will guarantee that all respondents are perceived to have $X = x_1$ with probability at least β. So, a device leading to a larger value of β will ensure greater privacy to respondents than one with a smaller β.

Let L, $0 < L < 1$, denote a preassigned level of respondents' privacy. Then in order to achieve this level of protection we require that $\beta \geq L$, irrespective of the values of $\pi_1, \ldots, \pi_m$. Thus we should have

$$(p\delta_{1j} + \frac{1-p}{m})\pi_1 \geq L(p\pi_j + \frac{1-p}{m}), \quad 1 \leq j \leq m,$$

or equivalently, the following inequalities should hold:

$$[p(1-L) + \frac{1-p}{m}]\pi_1 \geq \frac{L(1-p)}{m} \tag{14}$$

$$\text{and} \quad \frac{1-p}{m}\pi_1 - Lp\pi_j \geq \frac{L(1-p)}{m}, \quad 2 \leq j \leq m. \tag{15}$$

Clearly, no p can satisfy (13) irrespective of $\pi_1, \ldots, \pi_m$ for any given L since (13) fails as $\pi_1 \to 0$. So we assume that $\pi_1 > 0$ and we also assume some prior knowl-

edge about a lower bound on π_1. This assumption is quite realistic because in most populations there will be an appreciable number of persons with a non-stigmatizing variable value and hence, a lower bound to the proportion of such stigma-free persons in the population will be available.

Thus, suppose we have prior knowledge that $\pi_1 \geq c$. We work with $L < c$. This is again realistic because if the only knowledge about π_1 is that $\pi_1 \geq c$, it is impractical to demand that $P(X = x_1|R = x_j) \geq L(\geq c)$ for all j. Now, the following theorem gives the value of the device parameter p which will guarantee the desired level of respondent protection L.

Theorem 2 *Let β be as in (13) and $\pi_1 \geq c$ for some known c. Then given a preassigned L, where $0 < L < c$, $\beta \geq L$ will hold, irrespective of the values of $\pi_1, \ldots, \pi_m$, if and only if $p \leq p_0$, where*

$$p_0 = \frac{\frac{c-L}{m}}{\frac{c-L}{m} + L(1-c)}. \tag{16}$$

Proof Since $\pi_1 \geq c$, it is clear that $\pi_j \leq 1 - c$ for $2 \leq j \leq m$ and we have

$$\left[p(1-L) + \frac{1-p}{m}\right]\pi_1 \geq \left[p(1-L) + \frac{1-p}{m}\right]c$$
$$\text{and} \quad \frac{1-p}{m}\pi_1 - Lp\pi_j \geq \frac{1-p}{m}c - Lp(1-c), \quad 2 \leq j \leq m.$$

As a result, (14) and (15) will hold, irrespective of the true values of $\pi_1(\geq c), \pi_2, \ldots, \pi_m$ iff

$$\left[p(1-L) + \frac{1-p}{m}\right]c \geq L\frac{1-p}{m} \tag{17}$$
$$\text{and} \quad \frac{1-p}{m}c - Lp(1-c) \geq L\frac{1-p}{m} \tag{18}$$

hold. Now, (17) reduces to

$$\left(p + \frac{1-p}{m}\right)c \geq L\left(cp + \frac{1-p}{m}\right)$$

which will always hold for every p since $L(cp + \frac{1-p}{m}) \leq L(p + \frac{1-p}{m}) < c(p + \frac{1-p}{m})$ as $L < c$ and $p + \frac{1-p}{m} > 0$. So, it is enough to only consider (18). Note that

$$(18) \Leftrightarrow \frac{c - cp}{m} - Lp(1-c) \geq \frac{L - Lp}{m}$$
$$\Leftrightarrow p \leq \frac{\frac{c-L}{m}}{\frac{c-L}{m} + L(1-c)} = p_0,$$

thus proving the theorem. □

Table 1 Values of p_0 for various m and L

m	L	p_0	m	L	p_0	m	L	p_0
3	0.1	0.1413	4	0.1	0.1099	5	0.1	0.0899
3	0.2	0.2941	4	0.2	0.2381	5	0.2	0.2000
3	0.3	0.4494	4	0.3	0.3797	5	0.3	0.3288
3	0.4	0.5970	4	0.4	0.5263	5	0.4	0.4706

*R*emark The above discussion can be extended to include the more general case where X has t non-stigmatizing values $x_1, \ldots, x_t$, say, while its remaining $m - t$ values are stigmatizing, $1 < t < m$. In that case too, it can be shown that p_0 takes the form as in Theorem 2, but now with

$$\beta = \min_{1 \le j \le m} P(X = x_1 \text{ or } x_2 \text{ or } \ldots x_t | R = x_j) \text{ and } \pi_1 + \ldots + \pi_t \ge c \text{ with } L < c.$$

5 Privacy protection together with efficiency in estimation

We now consider the issue of efficiency in estimation together with privacy protection in randomized response surveys. It was seen from (5) that, irrespective of the values of $\pi_1, \ldots, \pi_m$, the efficiency of estimation may be increased by increasing p. On the other hand, for a given L and irrespective of the values of $\pi_1, \ldots, \pi_m$, Theorems 1 and 2 show that a protection of level L may be guaranteed iff $p \le p_0$, where p_0 is as in (8) or (16), respectively. So, the best choice of p with regard to maximizing the efficiency of estimation of μ_X, subject to the stipulated level of privacy protection L, is $p = p_0$. If we use a randomization device with p equal to any value *less than* p_0, then the efficiency of estimation will be less than that with $p = p_0$, even though the level of protection will still be L. The following examples illustrate this.

Example 5.1 Let X take four values which are all sensitive. Suppose $L = 0.1$. Then by Theorem 1, taking $m = 4$, we get $p_0 = 0.1099$. So, if we use a randomization device with $p = 0.1099$ then the efficiency of estimation can be maximized while guaranteeing that the maximum discrepancy between the true probability and the revealing probability of all respondents will be at most 0.1. However, if we use a device with $p > p_0$, then the desired level of privacy protection will not be realized.

Table 1 gives the p_0 values in (8) for some choices of L and m for achieving maximum efficiency of estimation.

For given m and L, if we use a device with $p < p_0$, then the efficiency of estimation will drop, even though the level of privacy protection will be guaranteed. To illustrate how the efficiency changes as p decreases from p_0, we use the case of $m = 3$, assuming for illustration that X takes the values $X = 1, 2, 3$ with probabilities 0.50, 0.35 and 0.25, respectively, in the population. In Table 2, for some illustrative values of p, we give the values of relative efficiency, defined as: $RelEff(p) = \{\text{Var}_{p_0}(\hat{\mu}_X)\}/\{\text{Var}_p(\hat{\mu}_X)\}$ where $\text{Var}_p(\hat{\mu}_X)$ is as given by (5). The p_0 values used are as given in Table 1.

Table 2 Relative Efficiencies for various values of p

L	p	$RelEff(p)$	L	p	$RelEff(p)$	L	p	$RelEff(p)$
0.1	0.141343	1	0.2	0.294118	1	0.3	0.449438	1
0.1	0.13	0.8452	0.2	0.27	0.8441	0.3	0.40	0..7847
0.1	0.12	0.7197	0.2	0.25	0.7184	0.3	0.36	0.6313
0.1	0.11	0.6043	0.2	0.23	0.6067	0.3	0.32	0.4957
0.1	0.10	0.4991	0.2	0.21	0.5047	0.3	0.28	0.3774
0.1	0.09	0.4040	0.2	0.19	0.4123	0.3	0.24	0.2759
0.1	0.08	0.3191	0.2	0.17	0.3295	0.3	0.20	0.1908
0.1	0.07	0.2889	0.2	0.15	0.2561	0.3	0.16	0.1217

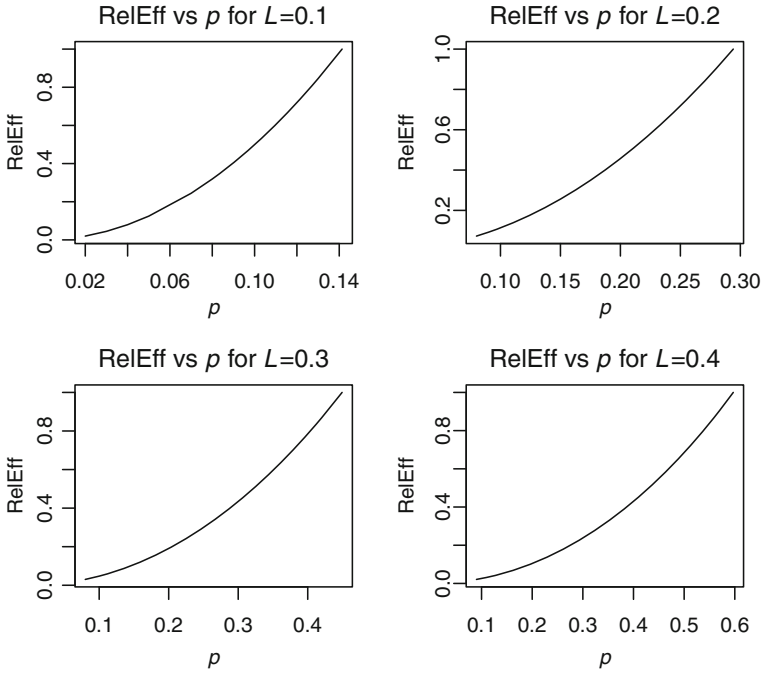

Fig. 1 Plot of Relative Efficiencies for various L values

Figure 1 shows how, for various values of L, the relative efficiencies drop as p is decreased from the optimal p_0 value.

Example 5.2 Let X take one nonsensitive value and two sensitive values. Suppose it can be assumed that at least 15% of the individuals in the population possess the nonsensitive value and suppose it is stipulated that $L = 0.10$. Then by Theorem 2, taking $m = 3$, $c = 0.15$, $L = 0.1$, we obtain $p_0 = 0.1639$. So, if we use a device with $p = 0.1639$ then estimation efficiency will be maximum while guaranteeing that all respondents will have at least a 10% probability of being revealed as belonging to the

non-stigmatizing class. As in Example 5.1, in this case too, the relative efficiencies drop as p is decreased from the optimal p_0 value.

6 Estimation of population proportions

As mentioned in Sect. 1, several researchers have estimated the proportions of individuals belonging to the two categories in dichotomous populations, while Loynes (1976) extended this to estimating the different proportions in a polychotomous population. In our case where X takes m numerical values, we may also readily estimate the population proportions $\pi_1, \ldots, \pi_m$ from the responses collected as in Sect. 2 and again use the measures of privacy as given in (7) and (13) to achieve the stipulated level of privacy protection.

As seen in Sect. 3, an unbiased estimate of π_i is

$$\hat{\pi}_i = \frac{1}{p}(w_i - \frac{1-p}{m}), \quad 1 \le i \le m.$$

Suppose, in the spirit of $A-$optimality commonly used in optimal design theory, we would like to minimize the average variance of these estimates. For this, we can show that the sum of the variances of the estimates of π_i is given by

$$\sum_{i=1}^{m} Var_p(\hat{\pi}_i) = \frac{1}{np^2}\sum_{i=1}^{m}\lambda_i(1-\lambda_i) = \frac{1}{n}\left\{\frac{1}{p^2} - \sum_{i=1}^{m}\pi_i^2 + \frac{1}{m}(\frac{1}{p^2} - 1)\right\}, \quad (19)$$

on simplification, using (4). Clearly, (19) is decreasing in p, irrespective of the true values of $\pi_1, \ldots, \pi_m$. So as in the case of estimating the mean, here too, subject to the stipulated level L of privacy protection, the best choice for p for minimizing the average variance of the estimates of the proportions may be obtained by applying Theorem 1 or 2, as the case may be. So, if all categories are sensitive, one uses $p = p_0$, with p_0 being given by (8) and if not all categories are sensitive, one uses p_0 given by (16). The following example illustrates this in the popular case of dichotomous populations.

Example 6.1 Suppose in a dichotomous population both categories are sensitive and we have to estimate the proportion of persons with these traits. Then the equivalent problem in our context is one where $m = 2$, i.e., X can take only two values x_1 and x_2 and we want to estimate the population proportions π_1 and π_2. This is because on the basis of values x_1 and x_2, the population units can be divided into 2 groups, say A and A^c.

When both A and A^c are sensitive, given L, one can apply Theorem 1 and compute p_0 using (8). For various levels of privacy protection as quantified by some illustrative values of L, the corresponding values of p_0 are given in Table 3, while the relative efficiency values for other values of p are shown in Table 4. When A is sensitive and A^c is not, we can proceed similarly by applying Theorem 2.

Figure 2 shows how the relative efficiencies drop as p is decreased from the optimal p_0 value.

Table 3 Values of p_0 for various L in a dichotomous population

L	0.1	0.2	0.3	0.4
p_0	0.1980	0.3846	0.5505	0.6897

Table 4 Relative Efficiencies for various values of p in a dichotomous population

L	p	$RelEff(p)$	L	p	$RelEff(p)$	L	p	$RelEff(p)$
0.1	0.1980	1	0.2	0.3846	1	0.3	0.5505	1
0.1	0.17	0.7317	0.2	0.36	0.8641	0.3	0.51	0.8298
0.1	0.15	0.5672	0.2	0.34	0.7629	0.3	0.47	0.6827
0.1	0.13	0.4244	0.2	0.32	0.6692	0.3	0.43	0.5555
0.1	0.10	0.2499	0.2	0.30	0.5829	0.3	0.39	0.4457
0.1	0.08	0.1595	0.2	0.28	0.5036	0.3	0.35	0.3512
			0.2	0.26	0.4308	0.3	0.31	0.2703
			0.2	0.24	0.3645	0.3	0.23	0.1444

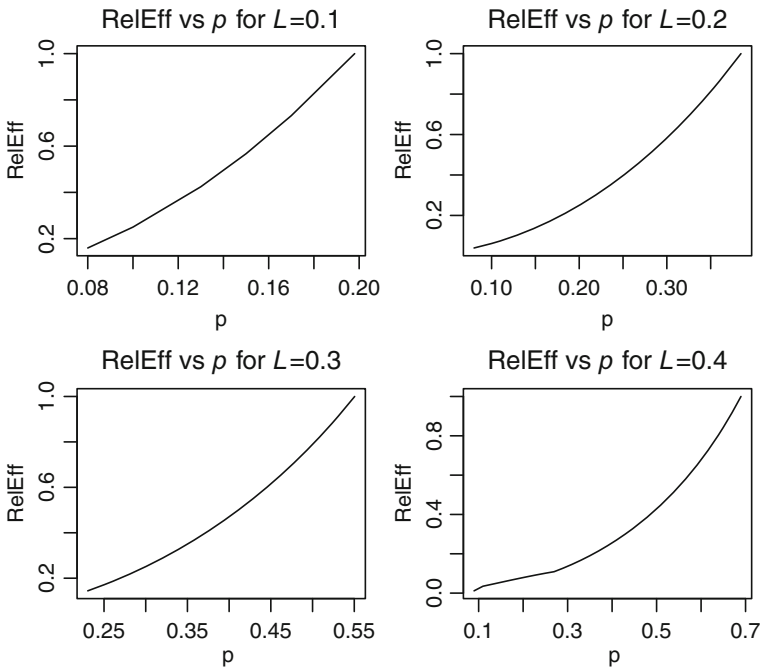

Fig. 2 Plot of Relative Efficiencies for dichotomous population

7 Concluding remarks

In this paper we have proposed a randomized response scheme for use in surveys where the sensitive or stigmatizing variable of interest is a discrete quantitative variable and

the target is to estimate the population mean. We focus our attention on the privacy protection afforded to respondents when they participate in a randomized response survey with this scheme and then develop measures of privacy protection. We study two broad situations: one where all values of the variable are sensitive and another where not all values are sensitive. In the latter case we elaborate on the case where only one of the possible values of the variable is non-stigmatizing whereas all remaining values of X are stigmatizing/sensitive and generalize to the case where t of the values of X are non-stigmatizing and the remaining values are not. We give examples to show that all these cases can arise in surveys.

We develop measures of privacy protection in these two situations and show that, given a target level of privacy protection, how the randomization device parameter may be chosen in order to achieve this level of protection. Finally we obtain the optimal value of the device parameter which allows the maximum efficiency of estimation while guaranteeing the desired level of privacy protection.

We show that our results may also be applied to ensure a desired level of privacy protection in the traditional studies with qualitative sensitive attributes in dichotomous (or polychotmous) populations where the target is to efficiently estimate the proportion (or proportions) of persons bearing the sensitive attribute (or attributes).

The issue of privacy protection when the sensitive variable is continuous and quantitative is yet to be developed. There is a need for a randomized response technique for such variables when the objective is to estimate the population mean efficiently while ensuring a given level of privacy protection. This problem is currently under investigation.

Acknowledgments The author is grateful to the reviewers for their careful reading of the earlier version and highly constructive comments.

References

Anderson H (1977) Efficiency versus protection in a general randomized response model. Scand J Stat 4:11–19

Arnab R, Dorffner G (2007) Randomized response techniques for complex survey designs. Stat Papers 48:131–141

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

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

Chaudhuri A, Bose M, Dihidar K (2011a) Estimation of a sensitive proportion by Warner's randomized response data through inverse sampling. Stat Papers 52:343–354

Chaudhuri A, Bose M, Dihidar K (2011b) Estimating sensitive proportions by Warner's randomized response technique using multiple randomized responses from distinct persons sampled. Stat Papers 52:111–124

Chaudhuri A, Mukerjee R (1987) Randomized response techniques: a review. Stat Neerlandica 41:27–44

Chaudhuri A, Mukerjee R (1988) Randomized responses: theory and techniques. Marcel Dekker, New York

Christofides TC (2005) Randomized response in stratified sampling. J Stat Plann Inference 128:303–310

Chua TC, Tsui AK (2000) Procuring honest responses indirectly. J Stat Plann Inf 90:107–116

Diana G, Perri PF (2009) Estimating a sensitive proportion through randomized response procedures based on auxiliary information. Stat Papers 50:661–672

Diana G, Perri PF (2011) A class of estimators for quantitative sensitive data. Stat Papers 52:633–650

Giordano S, Perri PF (2012) Efficiency comparison of unrelated question models based on same privacy protection degree. Stat Papers 53:987–999

Kim J (2007) A stratified unrelated question randomized response model. Stat Papers 48:215–233
Kuk AYC (1990) Asking sensitive questions indirectly. Biometrika 77:436–438
Lanke J (1975) On the choice of the unrelated question in Simmons' version of randomized response. J Amer Stat Assoc 70:80–83
Lanke J (1976) On the degree of protection in randomized interviews. Int Stat Rev 44:197–203
Leysieffer RW, Warner SL (1976) Respondent jeopardy and optimal designs in randomized response models. J Amer Stat Assoc 71:649–656
Ljungqvist L (1993) A unified approach to measures of privacy protection in randomized response models: a utilitarian perspective. J Amer Stat Assoc 88:97–103
Loynes RM (1976) Asymptotically optimal randomized response procedures. J Amer Stat Assoc 71:924–928
Mangat NS (1994) An improved randomized response strategy. J Roy Stat Soc 56:93–95
Nayak TK, Adeshiyan SA (2009) A unified framework for analysis and comparison of randomized response surveys of binary characteristics. J Stat Plann Inf 139:2757–2766
Pal S (2008) Unbiasedly estimating the total of a stigmatizing variable from a complex survey on permitting options for direct or randomized responses. Stat Papers 49:157–164
Van den Hout A, Van der Heijden PGM (2002) Randomized response, statistical disclosure control and misclassification: a review. Internat Stat Rev 70:269–288
Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. J Amer Stat Assoc 60:63–69