Stat Papers (2009) 50:661–672
DOI 10.1007/s00362-007-0107-y

NOTE

Estimating a sensitive proportion through randomized response procedures based on auxiliary information

Giancarlo Diana · Pier Francesco Perri

Received: 3 July 2007 / Revised: 9 November 2007 / Published online: 8 December 2007

Abstract Randomized response techniques are widely employed in surveys dealing with sensitive questions to ensure interviewee anonymity and reduce nonrespondents rates and biased responses. Since Warner's (J Am Stat Assoc 60:63–69, 1965) pioneering work, many ingenious devices have been suggested to increase respondent's privacy protection and to better estimate the proportion of people, π_A, bearing a sensitive attribute. In spite of the massive use of auxiliary information in the estimation of non-sensitive parameters, very few attempts have been made to improve randomization strategy performance when auxiliary variables are available. Moving from Zaizai's (Model Assist Stat Appl 1:125–130, 2006) recent work, in this paper we provide a class of estimators for π_A, for a generic randomization scheme, when the mean of a supplementary non-sensitive variable is known. The minimum attainable variance bound of the class is obtained and the best estimator is also identified. We prove that the best estimator acts as a regression-type estimator which is at least as efficient as the corresponding estimator evaluated without allowing for the auxiliary variable. The general results are then applied to Warner and Simmons' model.

Keywords Sensitive questions · Class of estimators · Regression estimator · Minimum variance bound

G. Diana
Department of Statistical Sciences, University of Padova,
Via Cesare Battisti 241, 35121 Padua, Italy
e-mail: giancarlo.diana@unipd.it

P. F. Perri (✉)
Department of Economics and Statistics, University of Calabria,
Via P. Bucci, 87036 Arcavacata di Rende (CS), Italy
e-mail: pierfrancesco.perri@unical.it

1 Introduction

In surveys concerning sensitive questions such as gambling, alcoholism, sexual behavior, drug taking, tax evasion, illegal income and else, direct techniques for collecting information may induce interviewed people to refuse answering or to give untruthful or misleading responses. To reduce nonrespondents rates and biased responses arising from sensitive, embarrassing, threatening, or even incriminating questions, some special statistical techniques may be employed to ensure interviewee anonymity or, at least, a higher degree of confidence. Such techniques, known as *randomized response methods*, use a randomization device, such as a die or a deck of cards, rather than a direct response to collect reliable information on sensitive issues. Depending on the result produced by the randomization device, the interviewee gives an answer concerning his/her true status. Since the interviewer is unaware of the result of the device, the use of these methods ensures that respondents can not be identified on the basis of their answers.

The first randomization method was introduced by Warner (1965) in order to gather trustworthy data to estimate π_A, the proportion of a population having a sensitive attribute, say A. Since Warner's pioneering work, many authors have dealt with the problem of estimating π_A and various methods have been suggested to make the technique more and more efficient. A concise description of some of these procedures may be found in the monographs by Fox and Tracy (1986), Chaudhuri and Mukerjee (1988), Hedayat and Sinha (1991) and Singh (2003). Among others, useful references on the subject are Mangat and Singh (1990), Mangat (1994), Tracy and Mangat (1996), Bhargava and Singh (2000), Chaudhuri (2004), Singh and Mathur (2002), Singh et al. (2003), Huang (2005), Shabbir and Gupta (2005), Saha (2006).

In sampling practice, direct techniques for collecting information about non-sensitive characters make massive use of auxiliary variables to improve sampling design and to achieve higher precision in population parameter estimates. Nevertheless, in spite of the broad variety of techniques developed to estimate non-sensitive parameters in the presence of auxiliary information, very few procedures have been suggested to improve randomization technique performance using supplementary information. Some attempts, developed in the context of sampling with unequal probabilities, are described in Chaudhuri and Mukerjee (1988), Allen and Singh (2001) and Grewal et al. (2006). Moreover, according to the simple random sampling, Zaizai (2006) used the auxiliary information directly at the estimation stage, improving Warner's estimator through the ratio method. Motivated by this work, we propose a general class of estimators which contains Warner and Zaizai's estimators and other estimators obtained from different randomization devices.

The remaining part of the paper is structured as follows. In Sect. 2, we briefly describe Warner's scheme together with its first alternative, known as the *unrelated-question model* or *Simmons' model*, and some further developments including the randomization device suggested by Zaizai. In Sect. 3, we introduce a general class of estimators for π_A, for a generic randomization scheme, and we obtain the minimum attainable variance bound for the class. The best estimator, which acts as a regression-type estimator, is also identified. It is pointed out that no further improvement over its efficiency is possible *ceteris paribus*. General results are then applied to Warner and Simmons' schemes. Section 4 concludes the work with some final considerations.

2 Warner's model and some variants

Let $P = \{1, 2, \ldots, N\}$ be a finite population of N individuals and Y a variable taking value $Y_i = 1$ if the ith person bears the sensitive characteristic A and $Y_i = 0$ otherwise, $i = 1, 2, \ldots, N$. To estimate the unknown proportion of people possessing the sensitive attribute, $\pi_A = N^{-1} \sum_{i=1}^{N} Y_i$, a sample of size n is selected according to the simple random sampling with replacement (*srswr)*. Henceforth, we will use capital letters for population and small ones for sample.

Warner (1965) suggested collecting information on the sensitive attribute by providing each respondent with a suitable randomization device, say a deck of cards. One of the two following statements is written on each card:

(i) *I possess the sensitive attribute A*
(ii) *I do not possess the sensitive attribute A*

in the proportion of p and $1 - p$, respectively.

The respondent randomly selects a card and gives a "yes" (or "no") response if his/her actual status matches (does not match) the statement on the card. Since the respondent does not reveal to the interviewer which card has been selected, his/her actual status can not be known by anyone.

To estimate π_A, assuming that the respondents were completely truthful in reporting their answers, Warner proposed the estimator

$$\hat{\pi}_W = \frac{\hat{\lambda} - (1 - p)}{2p - 1}, \qquad (p \neq 0.5) \tag{1}$$

where $\hat{\lambda} = (2p - 1)\bar{y} + (1 - p)$ is an unbiased estimator of $\lambda = (2p - 1)\pi_A + (1 - p)$, the proportion of "yes" answers in the population, $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$. The estimator is unbiased with variance

$$Var(\hat{\pi}_W) = \frac{\pi_A(1 - \pi_A)}{n} + \frac{p(1 - p)}{n(2p - 1)^2} \tag{2}$$

that can be unbiasedly estimated by

$$\hat{v}(\hat{\pi}_W) = \frac{\hat{\lambda}(1 - \hat{\lambda})}{(n - 1)(2p - 1)^2}. \tag{3}$$

A first alternative to Warner's method, aimed at enhancing the respondent's cooperation, is known as *Simmons' model*. The technique, firstly suggested by Horvitz et al. (1967) and then developed in Greenberg et al. (1969), allows the respondent to answer one of two questions in which one is concerned with the sensitive characteristic and the other is completely innocuous and unrelated to the sensitive attribute. Let Z be the non-sensitive variable taking values $Z_i = 1$ if the ith person possesses the innocuous attribute B, $Z_i = 0$ otherwise. In this way, Z and Y are uncorrelated.

To estimate π_A, each sampled respondent is provided with a randomization device consisting of two statements:

(i) *I possess the sensitive attribute A*
(ii) *I possess the innocuous attribute B*

which occur in the proportion of p and $1-p$, respectively. Analogously to Warner's approach, each respondent is asked to randomly select a card and report "yes" if his/her status matches the statement and "no" in the opposite situation.

Simmons' estimator of π_A is defined as

$$\hat{\pi}_S = \frac{\hat{\lambda} - (1-p)\pi_B}{p}, \qquad (p > 0)$$

where $\hat{\lambda} = p\bar{y} + (1-p)\bar{z}$ is again the unbiased estimator of λ, $\bar{z} = n^{-1}\sum_{i=1}^{n} z_i$.

We underline that λ ($\hat{\lambda}$) always denotes the proportion of "yes" answers in the population (sample) whose definition depends on the particular randomization device.

It is easy to prove that $\hat{\pi}_S$ is unbiased for π_A and has variance

$$\mathrm{Var}(\hat{\pi}_S) = \frac{\pi_A(1-\pi_A)}{n} + \frac{\pi_A(1-p)(1-2\pi_B)}{np} + \frac{\pi_B(1-p)[1-(1-p)\pi_B]}{np^2}. \tag{4}$$

An unbiased sample estimate of this variance is

$$\hat{v}(\hat{\pi}_S) = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)p^2}. \tag{5}$$

Different variants of Warner and Simmons' models have been proposed in the literature. Some of these rely on a randomization device with three outcomes. For instance, in addition to Warner's device, Mangat et al. (1995) suggested to add a blank card. The different cards occur in the proportions p_1, p_2 and p_3, $p_1 + p_2 + p_3 = 1$. If a blank card is selected, the respondent will report "no", irrespective of his/her actual status. Bhargava and Singh (2000) suggested a device that works in the same manner as the previous one, except for a slight difference: if a blank card is selected, the respondent will provide a "yes" response. In order to avoid that the blank cards are handled arbitrarily, Shabbir and Gupta (2005) required that, when a blank card appears, the respondent "speaks the truth" instead of saying "yes" or "no". Finally, Singh et al. (2003) adopted the procedure proposed by Mangat et al. (1995) to modify Simmons' model.

2.1 Randomized response strategies and auxiliary information

The conventional randomization methods are generally oriented toward the definition of a good compromise between respondents' privacy protection and precision of the estimates. In sampling surveys concerning non-sensitive issues, auxiliary information is commonly used to achieve higher precision in the estimates. This information may be used at the design stage of a survey yielding, for example, to stratification, systematic or unequal probability sampling designs, or directly at the estimation stage through the *ratio*, *product* or *regression methods*, or at both phases. On the contrary, auxiliary information has been scarcely employed to improve randomization strategy performance. Apart from the works dealing with the use of an auxiliary variable at the design stage as mentioned in Sect. 1, we recall Zaizai's (2006) recent work which integrated Warner's procedure providing an estimator of π_A based on the *ratio estimator* of the proportion of "yes" answers in the population.

Let X be a non-sensitive auxiliary variable, easy to collect, seemingly neutral but statistically connected with Y. Suppose that its mean, $\bar{X} = N^{-1}\sum_{i=1}^{N} X_i$, is known. Variables of this type are not unusual in many areas of social, clinical and medical research. For instance, on studying people's income and tax evasion, we may survey some variables such as the type of car, house size and people's neighborhood, which are certainly non-sensitive but connected with the people's living standards.

Under Warner's model, and assuming a positive correlation between Y and X, Zaizai (2006) suggested a randomization device consisting of a deck of cards showing the following two statements:

(i) *I possess the sensitive attribute A*
I possess the value X_i
(ii) *I do not possess the sensitive attribute A*
I possess the value X_i

in the proportion of p and $1 - p$. Once a card is selected, the respondent provides a "yes" (or "no") response if his/her actual status matches (does not match) the sensitive statement on the card and discloses the true value of the auxiliary variable.

We observe that only the response on the sensitive variable is randomized, whereas the value of X does not depend on the selected card since the interviewee is asked to always disclose the true value of the auxiliary variable.

According to the ratio method of estimation, π_A is estimated by

$$\hat{\pi}_Z = \frac{\hat{\lambda}_r - (1-p)}{2p-1} \tag{6}$$

where $\hat{\lambda}_r = \hat{\lambda}\bar{X}/\bar{x}$ is the ratio estimator of λ. For large samples, the estimator is unbiased and the variance is expressed as

$$\mathrm{Var}(\hat{\pi}_Z) = \mathrm{Var}(\hat{\pi}_W) - \frac{\lambda}{n(2p-1)^2}\left[2(2p-1)\pi_A C_{xy} - \lambda C_x^2\right] \tag{7}$$

where $C_{xy} = \sigma_{xy}/(\bar{X}\pi_A)$ and $C_x = \sigma_x^2/\bar{X}$; σ_{xy} and σ_x^2 denote the covariance between Y and X and the variance of X, respectively. Under suitable conditions, the estimator $\hat{\pi}_Z$ is demonstrated to be more efficient than Warner's one.

Table 1 Estimators belonging to the class $\hat{\pi}_g$

Estimator	b	c	h
Warner (1965)	0	$1-p$	$2p-1$
Greenberg et al. (1969)	0	$(1-p)\pi_B$	p
Mangat et al. (1995)	0	p_2	p_1-p_2
Bhargava and Singh (2000)	0	p_2+p_3	p_1-p_2
Huang (2005)	0	$1/2$	$1/2$
	0	1	$-1/2$
Shabbir and Gupta (2005)	0	p_2	$p_1-p_2+p_3$
Singh et al. (2003)	0	$p_2\pi_B$	p_1
Zaizai (2006)	$\hat{\lambda}/\bar{x}$	$1-p$	$2p-1$
Diana and Perri (2007a)	$(2p-1)\sigma_{xy}/\sigma_x^2$	$1-p$	$2p-1$

3 A general class of estimators

Let us consider a generic randomization device which may be, for instance, one of those mentioned in Sect. 2 or even a different one. According to the device, we propose a general class of estimators for π_A which employs an auxiliary variable X positively, as in Zaizai (2006), or negatively correlated with the sensitive one. Again, let us suppose that a sample of n respondents is selected according to *srswr*. Regarding the considered randomization device, each respondent is instructed to provide a randomized "yes" or "no" answer and to disclose the true value of the auxiliary variable. Under this approach, that generalizes Warner and Simmons' models, as well as their possible variants, we introduce the following class

$$\hat{\pi}_g = \frac{\hat{\lambda}_d - c}{h}, \quad (h \neq 0) \tag{8}$$

where $\hat{\lambda}_d = \hat{\lambda} - b(\bar{x} - \bar{X})$ is the *difference estimator* (see, e.g., Sukhatme et al. 1984); $\hat{\lambda}$ denotes a possible unbiased estimator of λ, and b, c, h are suitably chosen real constants. In particular, the values of c and h exclusively depend on the adopted randomization device, whereas the choice of b is mainly linked to the efficient use of the auxiliary variable. Therefore, in formulation (8), we can include both classical estimators without using supplementary information and possible estimators based on auxiliary variables. Some examples are given in Table 1.

Aimed at investigating the bias and the variance of the class, we observe that the precision of the estimates depends on both the randomization device and the sampling design. Therefore, we need to introduce some preliminary notation. Without loss of generality, let U be an indicator variable taking value 1 or 0 according to a "yes" or "no" response. The outcomes of U are determined by the randomization scheme chosen. Therefore, let $\lambda = N^{-1}\sum_{i=1}^{N} U_i$ be the unknown proportion of "yes" responses in the population and $\hat{\lambda} = \bar{u} = n_1/n$, where n_1 denotes the number of "yes" sample responses. It is easy to verify that the random variable n_1 follows a binomial distribution, $n_1 \sim Bin(n, \lambda)$.

Regardless of any randomization device, we now derive the exact expressions for the variance of the proposed class and for its unbiased estimator.

Proposition 1 *Under the srswr, if b is known and the constants c and h satisfy the constrain $c + h\pi_A = \lambda$, then $\hat{\pi}_g$ is unbiased for π_A with variance given by*

$$\mathrm{Var}(\hat{\pi}_g) = \frac{\lambda(1-\lambda)}{nh^2} + \frac{b}{nh^2}\left(b\sigma_x^2 - 2\sigma_{ux}\right) \tag{9}$$

which can be unbiasedly estimated using

$$\hat{v}(\hat{\pi}_g) = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)h^2} + \frac{b}{nh^2}\left(bs_x^2 - 2s_{ux}\right) \tag{10}$$

where $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ and $s_{ux} = \frac{1}{n-1}\sum_{i=1}^n (u_i - \bar{u})(x_i - \bar{x})$.

Proof The proof of (9) is straightforward and therefore is omitted. As for $\hat{v}(\hat{\pi}_g)$, we observe that $E(\hat{\lambda}) = \lambda$, $\mathrm{Var}(\hat{\lambda}) = \lambda(1-\lambda)/n$, while s_x^2 and s_{ux} are unbiased estimators of σ_x^2 and σ_{ux}, respectively. □

We notice that σ_{ux} depends on the adopted randomization device and the variance of $\hat{\pi}_g$ is the outcome of two components: the first is the variance of the corresponding estimator, say $\hat{\pi}^*$, evaluated without the auxiliary variable (i.e. $b = 0$), while the second refers to the variability introduced through the auxiliary information.

For a given sample design and for a fixed randomization device, the variance of $\hat{\pi}_g$ depends on the constant b. Among the different possible values for b, we chose b as the optimum solution of the minimization variance problem

$$b_o = \arg \min_b \mathrm{Var}(\hat{\pi}_g).$$

It is easy to verify that $\mathrm{Var}(\hat{\pi}_g)$ attains its minimum for

$$b_o = \frac{\sigma_{ux}}{\sigma_x^2} \tag{11}$$

which is the population linear regression coefficient of variable U on X.

Replacing b_o in (8), we obtain the best estimator in the class, say $\hat{\pi}_{g,\mathrm{opt}}$, which turns out to be a *regression-type estimator*. It attains the minimum variance bound of the class

$$\mathrm{Var}(\hat{\pi}_g)_{\min} = \frac{\lambda(1-\lambda)}{nh^2} - \frac{1}{nh^2}\frac{\sigma_{ux}^2}{\sigma_x^2}, \tag{12}$$

which, as in Cochran (1977, p. 195), can be estimated for large samples with

$$\hat{v}(\hat{\pi}_g)_{\min} = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)h^2} - \frac{1}{nh^2}\frac{s_{ux}^2}{s_x^2}. \tag{13}$$

An interesting result follows immediately from (12). Since $\text{Var}(\hat{\pi}^*) = \lambda(1-\lambda)/nh^2$ and $\sigma_{ux}^2/\sigma_x^2 \geq 0$, the best estimator in the class is at least as efficient as the estimator $\hat{\pi}^*$ defined without taking into account the auxiliary variable.

Unbiasedness, variance and variance estimation are now discussed under Warner and Simmons' schemes. Our findings are given in the following theorems.

Theorem 1 *Under Warner's randomization device, $\hat{\pi}_g$ is unbiased for π_A and has variance*

$$\text{Var}(\hat{\pi}_{g,w}) = \text{Var}(\hat{\pi}_W) - \frac{b}{n(2p-1)^2}\left[2(2p-1)\sigma_{xy} - b\sigma_x^2\right]. \tag{14}$$

Proof Considering the same randomization device described for Zaizai's procedure, we have: $U = pY + (1-p)(1-Y)$, $\lambda = p\pi_A + (1-p)(1-\pi_A)$, $c = 1-p$, $h = 2p-1$. Thus, from Proposition 1, $\hat{\pi}_{g,w}$ is unbiased. Moreover,

$$\sigma_{ux} = Cov\,[(2p-1)Y + (1-p), X] = (2p-1)\sigma_{xy}.$$

On substituting the previous expressions in (9) we get (14). Hence the proof. □

Theorem 2 *Under Simmons' randomization device, $\hat{\pi}_g$ is unbiased for $\hat{\pi}_A$ and has variance*

$$\text{Var}(\hat{\pi}_{g,s}) = \text{Var}(\hat{\pi}_S) - \frac{b}{np^2}\left(2p\sigma_{xy} - b\sigma_x^2\right). \tag{15}$$

Proof We consider two types of cards showing the following pairs of statements:

(i) *I possess the sensitive attribute A*
I possess the value X_i
(ii) *I possess the innocuous attribute B*
I possess the value X_i

represented with probability p and $1-p$, respectively. Unbiasedness and variance of $\hat{\pi}_{g,s}$ follow from Proposition 1 observing that now $U = pY + (1-p)Z$, $\lambda = p\pi_A + (1-p)\pi_B$, $c = (1-p)\pi_B$, $h = p$ and $\sigma_{ux} = p\sigma_{xy} + (1-p)\sigma_{xz}$, with $\sigma_{xz} = 0$ since $\sigma_{zy} = 0$ and $\sigma_{xy} \neq 0$. □

Corollary 1 *The variances of $\hat{\pi}_{g,w}$ and $\hat{\pi}_{g,s}$ are minimized for $b_o = (2p-1)\sigma_{xy}/\sigma_x^2$ and $b_o = p\sigma_{xy}/\sigma_x^2$, respectively, and are given by*

$$\text{Var}(\hat{\pi}_{g,w})_{\min} = \text{Var}(\hat{\pi}_W) - \frac{\sigma_{xy}^2}{n\sigma_x^2}, \quad \text{Var}(\hat{\pi}_{g,s})_{\min} = \text{Var}(\hat{\pi}_S) - \frac{\sigma_{xy}^2}{n\sigma_x^2}. \tag{16}$$

Sample estimates of $\mathrm{Var}(\hat{\pi}_{g,w})_{\min}$ *and* $\mathrm{Var}(\hat{\pi}_{g,s})_{\min}$*, valid in large samples, are*

$$\hat{v}(\hat{\pi}_{g,w})_{\min} = \hat{v}(\hat{\pi}_W) - \frac{s_{xy}^2}{ns_x^2}, \quad \hat{v}(\hat{\pi}_{g,s})_{\min} = \hat{v}(\hat{\pi}_S) - \frac{s_{xy}^2}{ns_x^2}. \tag{17}$$

Proof The result is straightforward and comes immediately from (3), (5), (11), (12) and (13). □

For Warner and Simmons' procedure, the use of the auxiliary variable X provides a regression-type estimator in the optimum case which is at least as efficient as the corresponding estimator obtained without employing the supplementary information. This result follows from (16) observing that

$$\mathrm{Var}(\hat{\pi}_W) \geq \mathrm{Var}(\hat{\pi}_{g,w})_{\min}, \quad \mathrm{Var}(\hat{\pi}_S) \geq \mathrm{Var}(\hat{\pi}_{g,s})_{\min}. \tag{18}$$

The estimators are equivalent only in case the variables Y and X are uncorrelated.

The results showed for Warner and Simmons' models may be easily adapted to any randomization device. Researchers need only to specify the expressions, induced by the randomization device, for U, λ, $\hat{\lambda}$, c, h and σ_{ux} and to determine the variance of the best estimator in the class. However, the regression-type estimator will always be at least as efficient as the corresponding estimator obtained under the randomization strategy which does not employ any auxiliary variable.

Remark The core of (8) relies on the attempt to better estimate λ through the *difference estimator*, $\hat{\lambda}_d$, which linearly combines the statistics $\hat{\lambda}$ and $\bar{x}$. For b known, this approach has the advantage of providing exact results for the unbiasness and the variance of the estimator of λ and of identifying the regression estimator as the more efficient one. As a matter of fact, many nonlinear different estimators of λ, based on the same amount of information, may be used following different ideas (see, e.g., Srivastava 1971). To the first order of approximation, i.e., up to terms which are $O(n^{-1})$, some of these estimators will be equivalent to the regression estimator of λ, while others will be less efficient. This result is not unexpected and may be justified adapting the approach discussed in Diana and Perri (2007b) where the regression estimator is found to be the best one for the unknown mean of a population when (multi-)auxiliary information is used. Consequently, since $\hat{\pi}_g$ is a linear transformation of $\hat{\lambda}_d$, no further improvement over the optimum $\hat{\pi}_{g,\mathrm{opt}}$ can be obtained to estimate π_A moving from nonlinear estimators of λ, at least to the first order of approximation.

3.1 The case of unknown b_o

In many situations, the optimum coefficient b_o defined in (11) is unknown. Usually, a good guess of b_o may be available on the basis of previous data, a pilot survey or past experience. However, if in (11) we replace a reliable guess, the variance of the estimator will not attain the minimum bound given in (12). If there are good reasons to believe that the guess is not reliable, then an estimate is to be calculated from the sample. A consistent estimate of b_o is likely to be the least squares estimate

$$\hat{b}_o = \frac{s_{ux}}{s_x^2} \tag{19}$$

which can always be calculated since u_i denotes the randomized response provided by the ith sampling respondent. We notice that this formulation avoids resorting the sensitive variable whose values can not be directly collected.

If we consider the estimate in (19), the results for b_o known are still valid to the first order of approximation on the basis of the well-known theory on the regression method of estimation (see, e.g., Sukhatme et al. 1984). This means that the estimator, say $\hat{\pi}_{\hat{g},\text{opt}}$, obtained by replacing the unknown b with $\hat{b}_o$ in $\hat{\pi}_g$, performs as well as the best estimator $\hat{\pi}_{g,\text{opt}}$. Each other estimator which employs an estimate of b different from $\hat{b}_o$ will not be optimum in the class. For instance, consider the estimator suggested by Zaizai (2006). Since $\hat{\pi}_Z$ is defined for $b = \hat{\lambda}/\bar{x}$, it is not optimum in the class and, thus, is less efficient than $\hat{\pi}_{g,\text{op}t}$ and $\hat{\pi}_{\hat{g},\text{opt}}$. Therefore, we can state that the use we make of the auxiliary variable through the regression method is certainly more profitable than the initial idea suggested by Zaizai. The gain in efficiency that derives from considering the best estimator instead of Zaizai's estimator may be quantified by the difference between $\text{Var}(\hat{\pi}_Z)$ and $\text{Var}(\hat{\pi}_{\hat{g},w})_{\min}$. To the first order of approximation, it is easy to verify that

$$\text{Var}(\hat{\pi}_Z) - \text{Var}(\hat{\pi}_{\hat{g},w})_{\min} \cong \frac{1}{n(2p-1)\sigma_x^2}\left[(2p-1)\sigma_{xy} - \lambda \bar{X} C_x^2\right]^2$$

which is always a non-negative quantity.

4 Conclusion

Traditional randomized response methods are based on devices which usually do not employ any auxiliary information for the estimation of the sensitive proportion π_A. Except for the first ideas described in Chaudhuri and Mukerjee (1988) and recently in Allen and Singh (2001), Grewal et al. (2006) and Zaizai (2006), no further development has been made to improve the precision of the estimates when auxiliary variables are available. Perhaps, this is due to the reluctance in providing responses which may be highly correlated with the sensitive variable and that may disclose respondents' privacy. Nevertheless, in social, clinical and medical surveys it may be possible to have one or more auxiliary variables available which are collected without additive cost for the survey and that show some association with the sensitive variable. For instance, researchers could consider social, economic or demographic information coming from administrative sources or which are in the public eye. In such circumstances, the auxiliary variables may be profitably used at the estimation stage, ensuring a high degree of privacy protection at the same time.

Aimed at seizing this opportunity we have proposed, in a general framework, a class of estimators for π_A involving a single auxiliary variable with known mean. Beside classical estimators for randomized responses, the class has the advantage of including the ratio-type estimator recently proposed by Zaizai (2006). Moreover, it may include many potential types of estimators which could be constructed employing the population and sample mean of the auxiliary variable.

For the estimators belonging to the proposed class, we have derived the minimum attainable variance bound and found the estimator which reaches this bound. In this optimum case, the proposed procedure provides an estimator of π_A which is at least as efficient as the analogous estimator evaluated without taking into account the auxiliary variable. Obviously, the gain in efficiency raises as the correlation between the auxiliary and the study variable increases.

The best estimator in the class acts as a regression-type estimator. Each estimator, based on the same amount of auxiliary information, may be at most as efficient as the best estimator. This aspect should avoid the proliferation of estimators apparently different from each other but that may perform at most as well as the regression estimator. More efficient estimators might be found using more than one auxiliary variable and/or integrating the information on the auxiliary means with that concerning, for instance, the variability and the shape of the variables.

References

Allen J, Singh S (2001) Response techniques to analyse various transformation and selection probabilities. Interstat. Available at http://interstat.statjournals.net/YEAR/2001/abstracts/0111002.php

Bhargava M, Singh R (2000) A modified randomization device for Warner's model. Statistica LX:315–321

Chaudhuri A (2004) Christofides'randomized response technique in complex sample survey. Metrika 60:223–228

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

Cochran WG (1977) Sampling techniques. Wiley, New York

Diana G, Perri PF (2007a) Regression type strategy for randomized response. In: Proceedings of the 2007 intermediate conference of the italian statistical society—risk and prediction—venice. 6-8 June, pp 459–460, Cleup, Padova

Diana G, Perri PF (2007b) Estimation of finite population mean using multi-auxiliary information. Metron LXV:99–112

Fox JA, Tracy PE (1986) Randomized response: a method for sensitive survey. Sage Publication, Inc., Newbury Park

Greenberg BG, Abul-Ela ALA, Simmons WR, Horvitz DG (1969) The unrelated question randomized response model: theoretical framework. J Am Stat Assoc 64:520–539

Grewal IS, Bansal ML, Sidhu SS (2006) Population mean corresponding to Horvitz–Thompson's estimator for multi-characteristics using randomized response technique. Model Assist Stat Appl 1:215–220

Hedayat AS, Sinha BK (1991) Design and inference in finite population sampling. Wiley, New York

Horvitz DG, Shah BV, Simmons WR (1967) The unrelated question randomized response model. In: Social statistics section proceedings of the American statistical association, pp 65–72

Huang K-C. (2005) Estimation of sensitive data from a dichotomous population. Stat Pap 47:149–156

Mangat NS (1994) An improved randomized response strategy. J R Stat Assoc B 56:93–95

Mangat NS, Singh R (1990) An alternative randomized response procedure. Biometrika 77:439–442

Mangat NS, Singh S, Singh R (1995) On use of a modified randomization device in Warner's model. J Indian Soc Stat Oper Res 16:65–69

Saha A (2006) A generalized two-stage randomized response procedure in complex sample survey. Aust NZ J Stat 48:429–443

Shabbir J, Gupta S (2005) On modified randomized device of Warner's model. Pak J Stat 21:123–129

Singh S (2003) Advanced sampling theory with applications, vol 2. Kluwer, Dordrecht

Singh HP, Mathur N (2002) On Mangat's improved randomized response strategy. Statistica LXII:397–403

Singh S, Horn S, Singh R, Mangat NS (2003) On the use of modified randomization device for estimating the prevalence of a sensitive attribute. Stat Transit 6:515–522

Srivastava SK (1971) A generalized estimator for the mean of a finite population using multi-auxiliary information. J Am Stat Assoc 66:404–407

Sukhatme PV, Sukhatme BV, Sukhatme S, Asok C (1984) Sampling theory of surveys with applications. Iowa State University Press, Ames

Tracy DS, Mangat NS (1996) On respondet's jeopardy in two alternative questions randomized response model. J Stat Plan Inference 55:107–114

Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc 60:63–69

Zaizai Y (2006) Ratio method of estimation of population proportion using randomized response technique. Model Assist Stat Appl 1:125–130