Metrika (2004) 60: 255–260
DOI 10.1007/s001840300308

Metrika

A note on randomized response models for quantitative data

Shaul K. Bar-Lev[1], Elizabeta Bobovitch and Benzion Boukai[2]

[1]Department of Statistics, University of Haifa Haifa 31905, Israel (E-mail: barev@stat.haifa.ac.il)
[2]Department of Mathematical Sciences, Indiana University, Purdue University, Indianapolis, Indiana 46202, USA

Abstract. Standard randomized response (RR) models deal primarily with surveys which usually require a 'yes' or a 'no' response to a sensitive question, or a choice for responses from a set of nominal categories. As opposed to that, Eichhorn and Hayre (1983) have considered survey models involving a quantitative response variable and proposed an RR technique for it. Such models are very useful in studies involving a measured response variable which is highly 'sensitive' in its nature. Eichhorn and Hayre obtained an unbiased estimate for the expectation of the quantitative response variable of interest. In this note we propose a procedure which uses a design parameter (controlled by the experimenter) that generalizes Eichhorn and Hayre's results. Such a procedure yields an estimate for the desired expectation which has a uniformly smaller variance.

1 Introduction

Since the pioneering work of Warner (1965) on randomized response (RR) models, various extensions have been introduced in the literature (see for example: Moors (1971), Chahudhuri (1987), Lakshmi and Raghvarao (1992), Mangat et al (1993)). The basic purpose of such methods and techniques is to estimate the proportion of a population whose truthful response to a sensitive question would be "yes", without exposing the respondents to the interviewer, and consequently avoiding social stigma or fear of reprisals. A good reference for some of the earlier works on the subject is the monograph by Chaudhuri and Mukerjee (1988). Recent publications on RR models are the works by Singh, Mangat and Singh (1997), Strachan, King and Singh (1998), Chua and Tsui (2000), Padmawar and Vijayan (2000), Chaudhuri (2001) and Gupta, Gupta and Singh (2002). Bar-Lev, Bobovich and Boukai (2003a) have proposed a two stage sequential sampling procedure for Warner's original

RR model, whereas a common Bayesian approach for several RR models has been suggested in Bar-Lev, Bobovich and Boukai (2003b).

Surveys employing RR techniques have been implemented in various fields. For example, Diskin and Felsenthal (1981) have analyzed the responses of Israeli interviewees with respect to sensitive issues relating to the Israeli society. In order to check the efficiency of the RR methodology, they have compared two sampling schemes; a simple random sample versus one obtained via an RR mechanism. Based on the results of the two samples as well as on corroborative external information obtained from objective sources (such as the Israeli Central Bureau of Statistics), they have strongly recommenced the use of RR techniques. Some other applications including a study of organized crime and a study of the incidents of abortions are discussed in Horvitz, Greenberg and Abernathy (1976).

The standard RR models deal with the question which usually requires yes or no (i.e., Bernoulli) response to the sensitive question, or allow a choice from a set of nominal categories (i.e., multinomial). As opposed to that, Eichhorn and Hayre (1983) have considered survey models involving a quantitative response variable and proposed an RR technique for it. Such models are very useful in studies involving a measured response variable which is highly 'sensitive' in its nature. For instance, studies addressing issues such as: a) Not only whether or not a woman had an abortion, but in addition, how many abortions she underwent; b) Not only whether the subject used illicit drugs, but also the number of occasions in which drugs were taken; c) Not only if an individual cheated on his income tax report, but also the amount of under reporting.

Eichhorn and Hayre obtained an unbiased estimate for the expected value of the quantitative response variable of interest and studied some of its immediate properties. In this note we propose a procedure which generalizes Eichhorn and Hayre's results and provide an alternative estimator to the mean response which has a uniformly smaller variance as compared to that of Eichhorn and Hayre (1983). For this reason, we briefly outline in Section 2, the Eichhorn and Hayre's estimation procedure. In Section 3 we present our extended model.

2 Eichhorn and Hayre's procedure

Eichhorn and Hayre (1983) considered an RR procedure appropriate for estimating the mean response when the sensitive variable of interest is quantitative in nature. By their procedure, the interviewees are asked about their value of the sensitive response variable. In return, they are allowed to respond with a coded (or scarmbled) value composed of their true value for the variable of interest, multiplied by some random number. The interviewer does not know which random number was used by each of the interviewees for coding their responses, but fully knows the underlying distribution which generated the random coding number.

Let X be a random variable (r.v.) denoting the quantitative response variable of interest and let Z be a r.v. representing the random number used in the coding mechanism. Assume that X (≥ 0) is independent of Z and let $Y = ZX$ be the coded response returned by the interviewee to the sensitive question. Also, denote

$$\mu_x = E(X),\ \mu_z = E(Z),\ \sigma^2 = V(X), \tau^2 = V(Z),$$

where μ_z and τ^2 are known and μ_x and σ^2 are unknown and write $c_x = \sigma/\mu_x$ and $c_z = \tau/\mu_z$ for the coefficient of variation of X and of Z, respectively. Then, it is straightforward to see that for the coded response, $Y = XZ$,

$$E(Y) = \mu_x\mu_z$$

and

$$V(Y) = \sigma^2\mu_z^2 + \mu_x^2(1 + c_x^2)\tau^2.$$

Based on a random sample $(Y_1, ..., Y_n)$ of coded responses of n interviewees, Eichhorn and Hayre proposed to estimate the unknown mean of the variable of interest, μ_x, by

$$\hat{\mu}_x = \frac{\bar{Y}}{\mu_z}\ ,$$

where $\bar{Y} = \sum Y_i/n$, is the sample mean of the n coded responses. It can easily be seen that $\hat{\mu}_x$ is an unbiased estimator of μ_x, with variance

$$V(\hat{\mu}_x) = \frac{1}{n}\left[\sigma^2 + \mu_x^2(1 + c_x^2)c_z^2\right],$$

which is larger than that resulting from a simple random sample with direct interviews; namely σ^2/n. Clearly if $P(Z = 1) = 1$, then the proposed technique turns out to be a direct interview, a fact which exposes the interviewee's response to the sensitive question. Accordingly, Eichhorn and Hayre have discussed different choices of Z for which $P(Z = 1) = 0$ as well as various alternatives for the distribution of Z so as to make the variance of $\hat{\mu}_x$ as small as possible.

3 The proposed procedure

In this section we propose a quantitative RR procedure which generalizes that of Eichhorn and Hayre and results in an estimate for μ_x whose variance is uniformly smaller than that of $\hat{\mu}_x$. This suggested procedure exploits both, the randomizing mechanism used in Warner's original RR model and the quantitative coding scheme in Eichhorn and Hayre (1983).

Let X and Z be as described above with $P(Z > 0) > 0$. Let $0 < p < 1$ be a design parameter, controlled by the experimenter, which is used for randomizing the interviewees' responses as follows: With probability p the interviewee responds with the true value of the quantitative variable X, whereas with probability $1 - p$ the interviewee responds with the coded variable ZX. That is, interviewee's responses to the sensitive question is,

$$Y = \begin{cases} X, & \text{with probability } p \\ ZX, & \text{with probability 1-}p. \end{cases}$$

Note that the design parameter p has a role similar to that used in Warner's model and that when $p = 0$ the proposed procedure reduces to that of Eichhorn and Hayre. Again, it is straightforward to see that the expectation and variance of the randomly coded response, Y, are given by

$$E(Y) = \mu_x(p + \mu_z(1 - p)) \tag{1}$$

and

$$V(Y) = \mu_x^2(1 + c_x^2)\left[p + \mu_z^2(1 + c_z^2)(1 - p)\right] - \mu_x^2(p + \mu_z(1 - p))^2. \tag{2}$$

Hence, the proposed estimate for μ_x, based on a random sample of the randomly coded responses, $Y_1, Y_2, \ldots, Y_n$ is

$$\hat{\mu}_x^* = \frac{\bar{Y}}{p + \mu_z(1 - p)} \ .$$

Clearly, by (1), $\hat{\mu}_x^*$ is an unbiased estimate for μ_x. By re-arranging the terms in (2), it follows that the expression for the variance of $\hat{\mu}_x^*$ can be written as,

$$V(\hat{\mu}_x^* \mid p) = \frac{1}{n}\left[\sigma^2 + \mu_x^2(1 + c_x^2)c_z^*(p)\right] \ , \tag{3}$$

where,

$$c_z^*(p) = \frac{p + E(z^2)(1 - p)}{(p + \mu_z(1 - p))^2} - 1.$$

Note that $c_z^*(p)$ is merely the squared coefficient of variation of the distribution of Z–but amended to include a point mass at 1 (with probability p). Now, for each given $0 < p < 1$, the variance of the randomly coded response Y in (2) may be estimated using the usual sample variance, $S_y^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2/(n - 1)$. Accordingly, for large n, the normal approximation can be utilized to provide a $(1 - \alpha)100\%$ confidence interval for μ_x as:

$$\hat{\mu}_x^* \pm \zeta_{1-\alpha/2}\frac{S_y}{\sqrt{n}(p + \mu_z(1 - p))},$$

where ζ_α denotes the $\alpha - th$ quantile of the standard normal distribution.

Note from (3), that the two values of $V(\hat{\mu}_x^* \mid p)$, namely,

$$V(\hat{\mu}_x^* \mid p = 1) = \sigma^2/n \ \text{ and } \ V(\hat{\mu}_x^* \mid p = 0) = V(\hat{\mu}_x) \tag{4}$$

represent, respectively, the two extreme situations: a direct response (interview) without any coding or scrambling mechanism and a situation in which all responses are coded, which is the procedure suggested by Eichhorn and Hayre. The next proposition shows that under mild conditions on the distribution of Z, and for all $p \in (0, 1)$, the variance of $\hat{\mu}_x^*$ is uniformly smaller than that of $\hat{\mu}_x$.

Proposition 1. *Assume that the distribution of Z satisfies*

$$0 < \mu_z < 2E(Z^2)/(1 + E(Z^2)). \tag{5}$$

Then, $\sigma^2/n < V(\hat{\mu}_x^* \mid p) < V(\hat{\mu}_x)$ for all $p \in (0, 1)$.

Proof. In view of (3) and (4), it is sufficient to show that the first derivative of $f(p) \doteq c_z^*(p)$, as a function of p, is negative for all $p \in (0, 1)$. Straightforward calculations show that

$$f(p) = \frac{(1 - E(Z^2))(p + \mu_z(1 - p)) - 2(1 - \mu_z)(p + E(Z^2)(1 - p))}{(p + \mu_z(1 - p))^3} \equiv \frac{C(p)}{D(p)}.$$

By assumption, $\mu_z > 0$ and therefore $D(p) > 0$ for all $p \in (0,1)$. Hence, we have to show that $C(p) < 0$, for all $p \in (0,1)$. To simplify the notation, let $a \equiv E(Z^2)$ and $b \equiv \mu_z$ and rewrite $C(p)$ in $f(p)$ as,

$$\begin{aligned} C(p) &\equiv (1-a)(p+b(1-p)) - 2(1-b)(p+a(1-p)) \\ &= p(a+b-ab-1) + (b-2a+ab), \end{aligned}$$

i.e., $C(p)$ is linear in p. By assumption (5), the constant term $b-2a+ab$ in $C(p)$ is negative. Consequently, if the coefficient of p in $C(p)$ is nonpositive then $C(p) < 0$ for all $p \in (0,1)$. Otherwise, if $a+b-ab-1 > 0$, then for all $p \in (0,1)$, $C(p) < C(1-) = 2b-(1+a)$. But since $b < 2a/(1+a) \leq (1+a)/2$ for all $a > 0$, it follows that $C(1-) < 0$ and hence $C(p) < 0$ for all $p \in (0,1)$. This completes the proof. ■

The proposed procedure for randomly coded quantitative response is easy to implement. As it is apparent from the results stated above, by an appropriate choice of the distribution of Z, this proposed procedure yields an estimate for the expectation of the quantitative (and 'sensitive') response variable which is uniformly more accurate, and thus more efficient, than the estimate suggested by Eichhorn and Hayre. Such distributions are easy to devise. For instance, the negative exponential distribution with mean $\mu_z = 1/\lambda$, where $2-\sqrt{2} < \lambda < 2+\sqrt{2}$, satisfies the assumption in (5) and would work well here.

For given σ^2 and μ_x, (so that c_x^2 is fixed), the relative efficiency of $\hat{\mu}_x^*$ to $\hat{\mu}_x$ can readily be seen from (3) and (4) to be,

$$eff^* = \frac{c_x^2 + (1+c_x^2)c_z^*(p)}{c_x^2 + (1+c_x^2)c_z^2}.$$

Similarly, the relative efficiency of the suggested estimator, $\hat{\mu}_x^*$, to that obtained without any coding of responses, namely by using $\bar{X}$, is

$$eff_0 = \frac{c_x^2 + (1+c_x^2)c_z^*(p)}{c_x^2}.$$

These efficiencies further illustrate the superiority of the suggested estimation procedure based on randomly coded responses, with $\hat{\mu}_x^*$ as compared to the estimate suggested by Eichhorn and Hayre. All in all, while the suggested procedure is less efficient than that of the direct response model, it offers at a cost of some efficiency a higher level of privacy protection to the interviewees and hence increasing their truthful cooperation with the survey, all at a higher accuracy compared to the fully coded response model suggested by Eichhorn and Hayre. However, the final choices of the design parameter p as well as the choice for the distribution of Z, and hence the choice of μ_z and τ, are to be determined by the accuracy considerations as well by the extent of interviewee's privacy protection that one feels is warranted in the given study. One should note, however, that further modifications of the current model, which are appropriate in other contexts, are possible and are straightforward to obtain.

Acknowledgement. We are grateful to two referees for their valuable and constructive comments.

References

[1] Bar-Lev SK, Bobovich E, Boukai B (2003a) A Two-Stage Sequential Scheme for Warner's Response Model. Communications in Statistics - Theory and Methods 32(12):2375–2389

[2] Bar-Lev SK, Bobovich E, Boukai B (2003b) A Common Conjugate Prior Structure For Several Randomized Response Models. Test 12:101–113

[3] Chaudhuri A (1987) Randomized Response Surveys of Finite Population: A Unified Approach for Quantitative Data. J of Statistical planning and Inference 15:157–165

[4] Chaudhuri A (2001) Using Randomized Response from a Complex Survey to Estimate a Sensitive Proportion in a Dichotomous Finite Population. J of Statistical Planning and Inference 94:37–42

[5] Chaudhuri A, Mukerjee R (1988) Randomized Response Theory and Techniques. Marcel Dekker, New-York

[6] Chua TC, Tsui AK (2000) Procuring Honest Responses Indirectly. J of Statistical Planning and Inference 90:107–116

[7] Diskin A, Felsenthal DS (1981) Do you lie? International Political Sciences Review 2:407–422

[8] Eichhorn BH, Hayre LS (1983) Scrambled Randomized Response Methods for Obtaining Sensitive Quantitative Data. J of Statistical planning and Inference 7:307–316

[9] Gupta S, Gupta B, Singh S (2002) Estimation of Sensitivity Level of Personal Interview Survey Questions. J of Statistical Planning and Inference 100(2):239–247

[10] Horvitz DG, Greenberg BG, Abernathy JR (1976) Randomized Response: A Data-Gathering Device for Sensitive Questions. International Statistical Review 44:181–186

[11] Lakshmi DV, Raghavarao D (1992) A Test for Detecting Untruthful Answering in Randomized Response Procedures. J of Statistical Planning and Inference 31:387–390

[12] Mangat NS, Singh R, Singh S, Singh B (1993) On Moors' Randomized Response Model. Biometrical Journal 35(6):727–732

[13] Moors JJA (1971) Optimization of the Unrelated Question Randomized Response Model. Jof the American Statistical Association 66: 627-629

[14] Padmawar VR, Vijayan K (2000) Randomized Response Revisited. J of Statistical Planning and Inference 90:293–304

[15] Singh S, Mangat NS, Singh R (1997) Estimation of Size and Mean of a Sensitive Quantitative Variable for a Sub-Group of a Population. Communications in Statistics - Theory and Methods 26:1793–1804

[16] Strachan R, King M, Singh S (1998) Likelihood-Based Estimation of the Regression Model with Scrambled Responses. Australian and New Zealand Journal of Statistics 40:279–290

[17] Warner SL (1965) Randomized Response : a Survey Technique for Elimination Evasive Answer Bias. J of the American Statistical Association 60:63–69