Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population

Arijit Chaudhuri *

Indian Statistical Institute, 203 Barrackpore, Trunk Road, Calcutta 700 035, India

Abstract

In estimating the proportion θ_A of people in a given community bearing a sensitive characteristic A, in order to protect the respondent's privacy, various techniques of generating randomized response (RR) rather than direct response are available in the literature. But the theory concerning them is developed only for samples selected by 'simple random sampling' (SRS) 'with replacement' or at most by SRS without replacement method. Illustrating two such RR devices we show how an estimator along with an estimated measure of its error may be developed if the sample of persons may be drawn adopting a complex survey design involving unequal selection probabilities with or without replacement.

MSC: 62 D05

Keywords: Dichotomous population; Estimation of a proportion; Randomized response; Sensitive characteristic; Unequal probabilities; With or without replacement sampling

1. Introduction

Warner (1965) introduced a technique of generating 'randomized response' (RR) as a device to protect a respondent's privacy when one intends to derive an unbiased estimator for the proportion θ_A of people bearing a sensitive characteristic, say, A, like habitual tax evasion, drunken driving, drug addiction, etc. Horvitz et al. (1967), Greenberg et al. (1969), Mangat and Singh (1990), Kuk (1990), Mangat (1992), Mangat et al. (1992), Moors (1997) among many others applied modifications on Warner's pioneering technique in order to promote respondent co-operation, improve upon the accuracy levels of the estimators and in general to advance the relevant theory and technique of RR. But one common limitation of these devices is that the sample in

* Tel.: +91-33-5778049/8085 Ex. 2816; fax: +91-33-577-6680.
E-mail address: achau@isical.ac.in (A. Chaudhuri).

each case must be drawn following the scheme of simple random sampling (SRS) with replacement (WR). In each case it is required to use the fact that while sampling on every draw from a given population of individuals, N in number, labeled $U=(1,\ldots,i,\ldots,N)$, say, 'the probability that a selected person bears A is $\theta_A=N_A/N$'; here N_A is the unknown number out of these N individuals who bear the stigmatizing characteristic A. If the sample is drawn with unequal selection probabilities 'with or without replacement' this 'probability' will be different and the methods of estimation have to be drastically revised. In what follows, we show how this may be achieved by illustrating with (1) Mangat's (1992) RR device and (2) another one with a modification thereupon. Similar modifications may be brought upon the other cases mentioned as well. But we omit details about them here.

2. Two specific RR devices and related estimation procedures

Mangat (1992) essentially considers the following RR device applying a modification on those by Warner (1965), Horvitz et al. (1967) and Greenberg et al. (1969). Suppose from a given population of N individuals of whom an unknown proportion θ_A bears A, an 'individual' is selected at 'random' and n such independent draws are made.

Each sampled person, following a given device, say, a table of random numbers or a pack of cards identical in all respects but differing only by certain 'distinguishing marks' is required, unnoticed by the interviewer, to (i) divulge truthfully, with a pre-assigned probability T $(0<T<1)$, whether he/she bears A or the complementary characteristic A^c i.e. whether the adopted random device 'matches' or 'mismatches' his/her true characteristic; and (ii) with a complementary probability $(1-T)$ to apply Horvitz et al.'s (1967) or Greenberg et al.'s (1969) device of giving out the RR in a way described below.

From a box of 'cards' marked either A or B in known proportions $p:1-p$ $(0<p<1)$, the sampled person is to randomly choose one card and truthfully divulge if the 'card mark' matches (or mismatches) his/her own true characteristic A or B (A^c or B^c – the corresponding complementaries). This is Mangat's (1990) 'two-stage' RR scheme.

Here B is a characteristic which need not be a stigmatizing one like A or even related in any way to it; for example, B may stand for 'vegetarian', 'flower lover', 'bespectacled' or some such innocuous features. This exercise is supposed to be done 'independently' by the individuals if sampled.

Mangat (1992) assumed that θ_B, the proportion of people in the community bearing the characteristic B, is 'known'. Then, he observed that the probability that a person chosen at 'random' from U may report that there is a 'match' between his/her true 'A or B' characteristic and the 'type' of card-mark is

$$\begin{aligned}\lambda &= T\theta_A+(1-T)[p\theta_A+(1-p)\theta_B]\\ &= C\theta_A+(1-T)(1-p)\theta_B,\end{aligned}$$

where $C = T + (1 - T)p$. This is true for every person in the SRSWR chosen on any of the n draws.

Assuming $C \neq 0$ and writing m for the number of 'matches' reported from the sample of n SRSWR draws, an unbiased estimator given by Mangat (1992) for θ_A is

$$\hat{\theta}_A = \left[\frac{m}{n} - (1 - T)(1 - p)\theta_B\right] \Big/ C. \tag{2.1}$$

Mangat (1992) also noted that its variance is

$$V(\hat{\theta}_A) = \frac{\lambda(1 - \lambda)}{nC^2}$$

and

$$v(\hat{\theta}_A) = \frac{(m/n)(1 - m/n)}{(n - 1)C^2},$$

is an unbiased estimator for $V(\hat{\theta}_A)$.

In order to develop estimators for θ_A along with variance or mean square error (MSE) estimators for them when the sample is chosen differently but the RR's by Mangat's two-stage scheme are used, let us proceed as below, first observing the following.

Let us write

$y_i = 1$ if a person labelled i bears A; 0 if i bears A^c,
$x_i = 1$ if a person labelled i bears B; 0 if i bears B^c,
$I_i = 1$ if a person labelled i announces 'match'; 0, else.

Then, $\theta_A = \sum y_i/N$, $\theta_B = \sum x_i/N$; and for Mangat's (1992) scheme of RR, no matter how a person labelled i is chosen, we have

$$\begin{aligned} Pr(I_i = 1) &= Ty_i + (1 - T)[py_i + (1 - p)x_i] = E_R(I_i) \\ &= Cy_i + d_i, \quad \text{where } d_i = (1 - T)(1 - p)x_i. \end{aligned}$$

Let $r_i = (I_i - d_i)/C$, assuming $C \neq 0$ and E_R, V_R the operators generically for expectation and variance with respect to any RR device.

$$\begin{aligned} \text{Then, } E_R(r_i) = y_i \quad \text{and} \quad V_R(r_i) &= E_R(I_i)(1 - E_R(I_i))/C^2 = V_i, \text{ say,} \\ &= a_i y_i + b_i, \text{ say, where} \end{aligned}$$

$$a_i = (1 - C - 2d_i)/C, \quad b_i = d_i(1 - d_i)/C^2$$

are known constants assuming x_i is known. Later we shall relax the requirement of x_i being known. It follows that $v_i = a_i r_i + b_i$ satisfies $E_R(v_i) = V_i$.

Let $Y = \sum y_i$; then $\theta_A = \bar{Y} = Y/N$. Let P denote any sampling design and $P(s)$ be the probability of selection of any sample s of distinct units from U. Let

$$t = t(s, \boldsymbol{Y}) = \sum y_i b_{si} I_{si},$$

be an estimator for Y with b_{si} chosen suitably as constants free of $\boldsymbol{Y} = (y_1, \ldots, y_i, \ldots, y_N)$, $I_{si} = 1$ if $i \in s$; 0 if $i \notin s$, in case y_i were available as DR's from the individuals i in s.

Then, with respect to the design P the expectation of t is $E_p(t)=\sum y_i E_p(b_{si}I_{si})$ and its MSE is $M_p(t)=E_p(t-Y)^2$ which becomes the variance $V_p(t)$ if $E_p(t)$ equals Y for every $\boldsymbol{Y}$ in which case t is unbiased for Y. Writing

$$M_p(t)=\sum y_i^2 C_i+\sum\sum_{i\neq j} y_i y_j C_{ij}$$

with C_i, C_{ij} determined by $b_{si}, P(s)$, let there exist an $m_p(t)=m_p(s,\boldsymbol{Y})=\sum y_i^2 d_{si}I_{si}+\sum\sum_{i\neq j} y_i y_j d_{sij}I_{sij}$ with $I_{sij}=I_{si}I_{sj}, d_{si}, d_{sij}$'s as constants free of $\boldsymbol{Y}$ such that

$$E_p(d_{si}I_{si})=C_i, E_p(d_{sij}I_{sij})=C_{ij}.$$

The literature has plenty of examples of such choices.

We shall restrict ourselves to the use of a P and a t such that $E_p(t)=Y$; then $m_p(t)$ is an unbiased variance estimator for t. Let us take

$$\begin{aligned} e&=\sum r_i b_{si}I_{si}\\ &=e(s,\boldsymbol{R}), \quad \text{where } \boldsymbol{R}=(r_1,\ldots,r_i,\ldots,r_N). \end{aligned} \tag{2.2}$$

Then, $E_R(e)=t$ and $E_p(e)=\sum r_i=R$, say. Then, the overall expectation of e is

$$E(e)=E_pE_R(e)=Y=E_RE_p(e), \quad \text{where } E=E_pE_R=E_RE_p.$$

In this sense e is an unbiased estimator for Y and hence $\bar{e}=e/N$ is a required unbiased estimator for $\bar{Y}=\theta_A$.

Then the variance of $\bar{e}$ is $V(\bar{e})=V(e)/N^2$, where V as the over-all variance operator. Now, $V(e)=E_pE_R(e-Y)^2=E_pE_R[(e-t)+(t-Y)]^2=E_p\sum V_i b_{si}^2 I_{si}+V_p(t)=E_RE_p[(e-R)+(R-Y)]^2=E_R[\sum r_i^2 C_i+\sum\sum_{i\neq j} r_i r_j C_{ij}]+\sum V_i$. Let $m_p(e)=\sum r_i^2 d_{si}I_{si}+\sum\sum_{i\neq j} r_i r_j d_{sij}I_{sij}$. Then, $E_p m_p(e)=\sum r_i^2 C_i+\sum\sum r_i r_j C_{ij}$ and $E_R m_p(e)=m_p(t)+\sum V_i d_{si}I_{si}$.

Then, it follows that

$$m(e)=m_p(e)+\sum v_i(b_{si}^2-d_{si})I_{si} \tag{2.3}$$

satisfies $Em(e)=V(e)$ and

$$m^*(e)=m_p(e)+\sum v_i b_{si}I_{si} \tag{2.4}$$

satisfies $Em^*(e)=V(e)$.

Then, $m(\bar{e})=m(e)/N^2$ and $m^*(\bar{e})=m^*(e)/N^2$ are two unbiased estimators for the variance of $\bar{e}$.

Let us next consider a modification of Mangat's (1992) RR technique when θ_B is unknown – this is a realistic situation in practice, though in the book Chaudhuri and Mukerjee (1988) there is a description about how a θ_B may be conveniently created.

In our procedure we just really need correspondingly a modification avoiding the requirement of the knowledge of x_i's. This is achieved in the following way.

Let everything else in Mangat's (1992) RR device remain the same except that instead of one there are two boxes. In the first box the A-marked and B-marked cards are mixed in proportions $p_1 : 1-p_1$ $(0<p_1<1)$. In the second box they are mixed in proportions $p_2 : 1-p_2$ $(0<p_2<1, p_1\neq p_2)$. Let each sampled person labelled i

be required to 'independently' draw 2 cards from both the boxes and truthfully divulge whether the 'card mark' matches his/her true characteristic A (or A^c) or B (or B^c). Let I_i, I_i' relate to the two draws from the first box and J_i, J_i' the same for the second box. Then

$$E_R(I_i) = Ty_i + (1-T)[p_1 y_i + (1-p_1)x_i] = E_R(I_i')$$

and

$$E_R(J_i) = Ty_i + (1-T)[p_2 y_i + (1-p_2)x_i] = E_R(J_i').$$

Let

$$r_i' = \frac{(1-p_2)I_i - (1-p_1)J_i}{(p_1-p_2)}, \quad r_i'' = \frac{(1-p_2)I_i' - (1-p_1)J_i'}{(p_1-p_2)}.$$

It follows that $E_R(r_i') = y_i = E_R(r_i'')$. Also,

$$V_R(r_i') = \frac{(1-p_2)^2 V_R(I_i) + (1-p_1)^2 V_R(J_i)}{(p_1-p_2)^2},$$

$$V_R(r_i'') = \frac{(1-p_2)^2 V_R(I_i') + (1-p_1)^2 V_R(J_i'')}{(p_1-p_2)^2}.$$

Now

$$V_R(I_i) = V_R(I_i') = E_R(I_i)(1 - E_R(I_i)),$$

$$V_R(J_i) = V_R(J_i') = E_R(J_i)(1 - E_R(J_i)).$$

So, $V_R(r_i') = V_R(r_i'') = W_i$, say.

Let $r_i = \dfrac{r_i' + r_i''}{2}$; then $E_R(r_i) = y_i$,

$$V_R(r_i) = \frac{W_i}{2} = V_i, \text{ say.}$$

Then,

$$v_i = \frac{(r_i' - r_i'')^2}{4} \text{ satisfies } E_R(v_i) = V_i.$$

So, with these choices of r_i and v_i used generically in (2.2)–(2.4) an unbiased estimator for θ_A along with two unbiased variance estimators follow immediately.

Since x_i need not be known here, it is this second modified RR technique that we recommend for application in survey sampling practice.

Acknowledgements

A referee's comments that led to this improved version of an earlier draft are gratefully appreciated.

References

Chaudhuri, A., Mukerjee, R., 1988. Randomized Response: Theory and Techniques. Marcel Dekker, New York.

Greenberg, B.G., Abul-Ela, E.L.A., Simmons, W.R., Horvitz, D.G., 1969. The unrelated question randomized response model: theoretical framework. J. Amer. Stat. Assoc. 64, 520–539.

Horvitz, D.G., Shah, B.V., Simmons, W.R., 1967. The unrelated question randomized response model. Proceedings of the Social Stat. Sec. Amer. Stat. Assoc. 65–72.

Kuk, A.Y.C., 1990. Asking sensitive question indirectly. Biometrika 77, 436–438.

Mangat, N.S., 1992. Two stage randomized response sampling procedure using unrelated question. J. Ind. Soc. Agri. Stat. 44 (1), 87–88.

Mangat, N.S., Singh, R., 1990. An alternative randomized response procedure. Biometrika 77, 439–442.

Mangat, N.S., Singh, R., Singh, S., 1992. An improved unrelated question randomized response strategy. Cal. Stat. Assoc. Bull. 42, 277–281.

Moors, J.J.A., 1997. A critical evaluation of Mangat's two-step procedure in randomized response. A discussion paper of Tillburg University.

Warner, S.L., 1965. RR: a survey technique for eliminating evasive answer bias. J. Amer. Stat. Assoc. 60, 63–69.