ELSEVIER

Journal of Statistical Planning and Inference 90 (2000) 293–304

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Randomized response revisited

V.R. Padmawar[a, *, 1], K. Vijayan[b]

[a] *Statistics and Mathematics Division, Indian Statistical Institute, 8th Mile, Mysore Road, Bangalore 560 059, India*

[b] *Department of Mathematics, The University of Western Australia, Nedlands, WA 6907, Western Australia, Australia*

Received 21 April 1999; received in revised form 10 July 1999; accepted 19 January 2000

Abstract

The randomized response techniques for collecting data on sensitive variables have been in vogue for over three decades now. In this note we introduce a method of randomization that hitherto has not been resorted to. We first develop a theory consequent to the new technique of randomization and later contrast it with the traditional one. © 2000 Elsevier Science B.V. All rights reserved.

MSC: primary 62D05

Keywords: Sensitive variable; Randomized response; Unbiased estimator; Admissibility; Best estimator; Superpopulation model

1. Introduction

In survey methodology whenever the study variable is sensitive in nature either because it pertains to something that is too personal or stigmatizing or illegal, randomized response (RR) techniques are used to collect the data. A typical RR method (see Chaudhuri and Mukerjee (1987), Hedayat and Sinha (1991)) may be described as a procedure in which the respondent reports either the study variate value y or some other innocuous variate value x with some prespecified probabilities. The Statistician, does not, however, know whether the reported value corresponds to the variable x or the variable y. This may be viewed as the Statistician receiving either the signal y or the noise x with some specified probabilities. The Statistician's objective is to estimate some function of the signals.

Instead, we may think of a method by way of which the Statistician receives a 'mixture of signal and noise'. In other words, the respondent reports a value which

* Corresponding author.

E-mail addresses: vrp@isibang.ac.in (V.R. Padmawar), vijayan@maths.uwa.edu.au (K. Vijayan).

[1] This work was done while visiting the University of Western Australia.

PII: S0378-3758(00)00097-5

is obtained by adding a value of random variable to the true Study variate value say, $y + x$. This may be viewed as signal plus noise. The Statistician has to 'extract' the signals to be able to carry out the estimation or inference.

As the respondent does not ever have to reveal the true study variate value the procedure allows complete anonymity to the respondent thereby ensuring absolute protection of privacy. Consequently, the respondent would feel extremely reassured to adhere to truthful reporting. For example, say the Statistician is interested in collecting data on y: the number of induced abortions. The respondent adds her true y-value to the randomly generated observation x from the normal distribution $N(-10, 1)$, say, and reports the value $y + x$ to the Statistician. If the respondent has to report a value -8.712, say, she would feel absolutely safe and extremely encouraged to report such a value as it does not seem to have any relationship, whatsoever, with the true number of induced abortions. Thus, the new method has a clear psychological edge over the traditional RR methods. We would later establish that the new method also has a theoretical advantage in the sense of being able to achieve reduction in the variability of the estimators.

2. Preliminaries

Let $U = \{1, 2, \ldots, N\}$ be a finite population under consideration. Let y be the sensitive study variate under consideration that takes value y_i on unit i, $1 \leqslant i \leqslant N$. It is required to estimate the population total $T(\boldsymbol{y}) = \sum_{i=1}^{N} y_i$ based on sample s of size n drawn using a given sampling design p.

Let X be a random variable with known distribution function F. In particular, let X have mean μ and variance σ^2.

Each of the selected individuals $i \in s$ would be asked to draw an observation X_i from F. The respondent would then be asked to report the value

$$z_i = y_i + x_i,$$

where y_i is the true y-value of the ith respondent, $i \in s$.

We assume that this reporting would be done truthfully and correctly.

The entire data set based on the sample s may now be represented as

$$\textbf{data}: \{z_i = y_i + x_i \mid i \in s\}. \tag{2.1}$$

With reference to the data set (2.1) it must be borne in mind that there are two sources of randomness. The first one is the design-based randomness generating different samples s according to the given sampling design p. The second source of randomness is the random observation from the distribution function F. The latter source of randomness is responsible for randomized response.

Assume for the time being that the true y_i-values can be had through direct response. Then a linear unbiased estimator for $T(\boldsymbol{y}) = \sum_{i=1}^{N} y_i$ is of the type

$$e(s, \boldsymbol{y}) = \sum_{i=1}^{N} b(s, i) y_i, \tag{2.2}$$

where the coefficients satisfy

$$b(s,i) = 0 \quad \text{if } i \notin s$$

and

$$E_p(b(s,i)) = 1 \quad \forall i = 1,2,\ldots,N. \tag{2.3}$$

The purpose of the RR methods is to provide a means to obtain estimators for $e(s,\boldsymbol{y}) = \sum_{i=1}^{N} b(s,i)y_i$, which itself is an estimator for $T(\boldsymbol{y}) = \sum_{i=1}^{N} y_i$, using the RR data set (2.1). In particular, we may think of obtaining an estimator for the true value of y_i, using the RR variable z_i attached to unit i. In what follows, we construct such estimators and study their properties.

3. Constructing unbiased estimators

Let $E_p, V_p, \ldots,$ stand for the design-based operations and E_R, $V_R, \ldots,$ stand for the operations based on the randomization method. Here it would mean the operations based on the distribution function F.

Now

$$z_i = y_i + x_i \Rightarrow E_R(z_i) = y_i + \mu.$$

Therefore,

$$\hat{y}_i = z_i - \mu, \tag{3.1}$$

serves as an unbiased estimator for y_i, $i \in s$, in the sense that

$$E_R(\hat{y}_i) = E_R(y_i + x_i - \mu) = y_i.$$

We are now in a position to prove the following theorem.

Theorem 3.1. *The estimator*

$$e(s,\hat{\boldsymbol{y}}) = \sum_{i=1}^{N} b(s,i)\hat{y}_i \tag{3.2}$$

is an unbiased estimator of the population $T(\boldsymbol{y})$ in the sense that

$$E_p E_R(e(s,\hat{\boldsymbol{y}})) = T(\boldsymbol{y}),$$

where $\hat{y}_i$, $i \in s$, are as in (3.1).

Proof. The proof is obvious. □

We now move on to compute the expression for the variance of the estimator $e(s,\hat{\boldsymbol{y}})$. Using the symbolic formula

$$V = V_1 E_2 + E_1 V_2,$$

we get

$$V(e(s,\hat{\boldsymbol{y}})) = V_p E_R(e(s,\hat{\boldsymbol{y}})) + E_p V_R(e(s,\hat{\boldsymbol{y}}))$$

$$= V_p(e(s,\boldsymbol{y})) + E_p \sum_{i=1}^{N} b^2(s,i) V_R(\hat{y}_i)$$

$$= V_p(e(s,\boldsymbol{y})) + \sigma^2 \sum_{i=1}^{N} E_p(b^2(s,i)) \quad (\text{as } V_R(\hat{y}_i) = \sigma^2).$$

Thus,

$$V(e(s,\hat{\boldsymbol{y}})) = V_p(e(s,\boldsymbol{y})) + \sigma^2 \sum_{i=1}^{N} E_p(b^2(s,i)). \tag{3.3}$$

The first term on the right-hand side is the design variance of the estimator $e(s,\boldsymbol{y})$ under direct response, the second term is the additional variability due to randomized response. For instance, $\sigma^2 = 0$ corresponds to degenerate X, hence to direct response, then as expected, $V(e(s,\hat{\boldsymbol{y}}))$ coincides with $V_p(e(s,\boldsymbol{y}))$.

We now obtain an estimator for $V(e(s,\hat{\boldsymbol{y}}))$.

First of all, note that $V_p(e(s,\boldsymbol{y}),\boldsymbol{y})$, the expression for the variance of the estimator $e(s,\boldsymbol{y})$ at the vector $\boldsymbol{y}$, is given by

$$V_p(e(s,\boldsymbol{y}),\boldsymbol{y}) = \sum_{i=1}^{N} a_{ii} y_i^2 + \sum_{i\neq j=1}^{N} a_{ij} y_i y_j, \tag{3.4}$$

where

$$a_{ii} = E_p(b^2(s,i)) - 1, \quad 1 \leqslant i \leqslant N,$$

$$a_{ij} = E_p(b(s,i)b(s,j)) - 1, \quad 1 \leqslant i \neq j \leqslant N.$$

Let

$$\hat{V}_p(e(s,\boldsymbol{y}),\boldsymbol{y}) = \sum_{i=1}^{N} a_{ii}(s) y_i^2 + \sum_{i\neq j=1}^{N} a_{ij}(s) y_i y_j, \tag{3.5}$$

where

$$a_{ii}(s) = 0 \quad \text{if } i \notin s,$$

$$a_{ij}(s) = 0 \quad \text{if } i \notin s \text{ or } j \notin s$$

and

$$E_p a_{ii}(s) = a_{ii}, \quad 1 \leqslant i \leqslant N,$$

$$E_p a_{ij}(s) = a_{ij}, \quad 1 \leqslant i \neq j \leqslant N.$$

Clearly, this is an unbiased estimator for $V_p(e(s,\boldsymbol{y}))$ under design p.

Let

$$\hat{V}_p(e(s,\boldsymbol{y}),\hat{\boldsymbol{y}}) = \sum_{i=1}^{N} a_{ii}(s) \hat{y}_i^2 + \sum_{i\neq j=1}^{N} a_{ij}(s) \hat{y}_i \hat{y}_j. \tag{3.6}$$

Note that (3.6) is the same as (3.5) with y_i's replaced by $\hat{y}_i$'s.

Observe that

$$E_p E_R \hat{V}_p(e(s,\boldsymbol{y}),\hat{\boldsymbol{y}}) = E_p\left\{\sum_{i=1}^{N} a_{ii}(s)(y_i^2+\sigma^2) + \sum_{i\neq j=1}^{N} a_{ij}(s)y_i y_j\right\}$$

$$= \sum_{i=1}^{N} a_{ii}y_i^2 + \sum_{i\neq j=1}^{N} a_{ij}y_i y_j + \sigma^2\sum_{i=1}^{N} a_{ii}.$$

Hence,

$$E_p E_R \hat{V}_p((e(s,\boldsymbol{y}),\hat{\boldsymbol{y}})) = V_p(e(s,\boldsymbol{y})) + \sigma^2\sum_{i=1}^{N} E_p(b^2(s,i)) - N\sigma^2. \tag{3.7}$$

We are now in a position to state the following theorem.

Theorem 3.2. *The variance* $V(e(s,\hat{\boldsymbol{y}}))$ *of* (3.3) *can be estimated unbiasedly by*

$$\hat{V}_p(e(s,\boldsymbol{y}),\hat{\boldsymbol{y}}) + N\sigma^2, \tag{3.8}$$

where $\hat{V}_p(e(s,\boldsymbol{y}),\hat{\boldsymbol{y}})$ *is given by* (3.6).

Proof. The proof is immediate using (3.7). □

$N\sigma^2$ in (3.8) may be replaced by suitable unbiased estimators of $N\sigma^2$ to get different unbiased estimators for $V(e(s,\hat{\boldsymbol{y}}))$.

There is yet another way of constructing an estimator for $T(\boldsymbol{y})$. Let

$$e(s,\boldsymbol{z}) = \sum_{i=1}^{N} b(s,i)z_i. \tag{3.9}$$

Now,

$$E_p E_R e(s,\boldsymbol{z}) = E_p\sum_{i=1}^{N} b(s,i)(y_i+\mu)$$

$$= \sum_{i=1}^{N} E_p b(s,i)(y_i+\mu)$$

$$= T(\boldsymbol{y}) + N\mu.$$

Hence, the estimator

$$e(s,\boldsymbol{z}) - N\mu \tag{3.10}$$

is unbiased for $T(\boldsymbol{y})$.

To compute the variability of the above estimator it is enough to look at the variability of $e(s,\boldsymbol{z})$ as $N\mu$ is a constant.

We again use the symbolic formula $V = V_1E_2 + E_1V_2$,

$$V(e(s,\boldsymbol{z})) = V_p E_R\sum_{i=1}^{N} b(s,i)z_i + E_p V_R\sum_{i=1}^{N} b(s,i)z_i$$

$$= V_p\sum_{i=1}^{N} b(s,i)(y_i+\mu) + \sigma^2 E_p\sum_{i=1}^{N} b^2(s,i).$$

Thus,

$$V(e(s,\boldsymbol{z})) = V_p(e(s,\boldsymbol{y}+\boldsymbol{\mu})) + \sigma^2 \sum_{i=1}^{N} E_p(b^2(s,i)), \tag{3.11}$$

where $V_p(e(s,\boldsymbol{y}+\boldsymbol{\mu}))$ is the design variance of $e(s,\boldsymbol{y})$ at the vector $\boldsymbol{y}+\boldsymbol{\mu}$, where $\boldsymbol{\mu}=\mu\mathbf{1}$ and $\mathbf{1}' = (1,1,\ldots,1)$.

To compare estimators (3.2) and (3.10) we compare expressions (3.3) and (3.11).

Theorem 3.3. *Estimator* (3.2) *is better than estimator* (3.10) *if and only if*

$$V_p(e(s,\boldsymbol{y})) \leqslant V_p(e(s,\boldsymbol{y}+\boldsymbol{\mu})).$$

Remark 3.1. In practice, if we start with an estimator $e(s,\boldsymbol{y})$ that is expected to perform well at $\boldsymbol{y}$ in the direct response set-up then it may be reasonable to expect that estimator (3.2) performs better than estimator (3.10) in the RR set-up.

The variance in (3.11) can be estimated as in the earlier case.

We have

$$\hat{V}_p(e(s,\boldsymbol{y}),\hat{\boldsymbol{y}}+\boldsymbol{\mu}) - N\sigma^2 \tag{3.12}$$

as an unbiased estimator for the variance in (3.11), where $\hat{V}_p(e(s,\boldsymbol{y}),\ \hat{\boldsymbol{y}}+\boldsymbol{\mu})$ is obtained by replacing $\hat{y}_i$ in (3.6) by $\hat{y}_i+\mu$.

Remark 3.2. In fact, we can think of a class of unbiased estimators for $T(\boldsymbol{y})$ obtained by using convex combinations of the two estimators (3.2) and (3.10).

Consider the following class of unbiased estimators indexed by a:

$$\left\{ e_a = e(s,\boldsymbol{z}) - \left(a + \frac{1-a}{N}\sum_{i=1}^{N} b(s,i)\right) N\mu \;\middle|\; 0 \leqslant a \leqslant 1 \right\}. \tag{3.13}$$

We again use the symbolic formula $V = V_1E_2 + E_1V_2$ to compute the variability of the estimator e_a,

$$V(e_a) = E_pV_R(e(s,\boldsymbol{z})) + V_p(e(s,\boldsymbol{y},\boldsymbol{y}+a\mu\mathbf{1})).$$

The second term on the right-hand side of the above expression depends on a. Let us try to minimize that. Let

$$Q(a) = (\boldsymbol{y}+a\mu\mathbf{1})'A(\boldsymbol{y}+a\mu\mathbf{1}),$$

where A is a nonnegative-definite matrix. Then,

$$\frac{\partial Q}{\partial a} = 2\mu\mathbf{1}'A(\boldsymbol{y}+a\mu\mathbf{1}),$$

$$\frac{\partial^2 Q}{\partial a^2} = 2\mu^2\mathbf{1}'A\mathbf{1}.$$

If $\mathbf{1}$ happens to be an eigenvector of the matrix A corresponding to the eigenvalue 0 then $Q(a)$ is independent of a.

Otherwise $\mathbf{1}'A\mathbf{1} > 0$, in which case, solving $\partial Q/\partial a = 0$ for optimal a^* we get

$$a^* = -\frac{\mathbf{1}'A\boldsymbol{y}}{\mu\mathbf{1}'A\mathbf{1}}.$$

It remains to be checked whether $0 \leqslant a^* \leqslant 1$.

It should, however, be noted that the optimal a^* depends on the unknown vector $\boldsymbol{y}$.

Remark 3.3. The practical significance of the above result is that if we start with a 'good' estimator e then a reasonable estimator of $\mathbf{1}'A\boldsymbol{y}$ may be used so that μ can suitably be chosen to get $0 \leqslant \hat{a^*} \leqslant 1$, and finally use the estimator $e_{\hat{a^*}}$.

4. Admissibility and UMVLUE

In this section we first prove a result pertaining to the admissibility of the derived estimator and then move on to prove the nonexistence of a best estimator. Here we assume that $\Theta \subset \mathbb{R}^N$, the parametric space for $\boldsymbol{y}$, is such that if $\boldsymbol{w} \in \Theta$ then $\lambda\boldsymbol{w} \in \Theta \ \forall \lambda > 0$.

Theorem 4.1. *For a given design p if $e(s,\boldsymbol{y})$ is an admissible linear unbiased estimator in the direct response set-up then so is the derived estimator $e(s,\hat{\boldsymbol{y}})$ in the RR set-up.*

Proof. Let the linear unbiased estimator $e(s,\boldsymbol{y}) = \sum_{i=1}^{N} b(s,i)y_i$ be admissible in the direct response set-up. Let further

$$B(p,e) = B(e) = B(e(s,\boldsymbol{y})) = \sum_{i=1}^{N} E_p(b^2(s,i)). \tag{4.1}$$

Note that the expression $B(e)$ is independent of $\boldsymbol{y}$.

Recall that from (3.3)

$$V(e(s,\hat{\boldsymbol{y}})) = V_p(e(s,\boldsymbol{y})) + \sigma^2 B(e). \tag{4.2}$$

If possible let there exist an estimator $e_1(s,\boldsymbol{y}) = \sum_{i=1}^{N} b_1(s,i)y_i$ such that the derived estimator $e_1(s,\hat{\boldsymbol{y}})$ is better than the derived estimator $e(s,\hat{\boldsymbol{y}})$ in the randomized response set-up.

Since the estimator $e(s,\boldsymbol{y})$ is admissible in the direct response set-up it would be better than the estimator $e_1(s,\boldsymbol{y})$ at some vector, say $\boldsymbol{y}_0$, i.e.,

$$V_p(e_1(s,\boldsymbol{y}_0)) > V_p(e(s,\boldsymbol{y}_0)).$$

Thus, choosing λ suitably we can make

$$V_p(e_1(s,\lambda \boldsymbol{y}_0)) - V_p(e(s,\lambda \boldsymbol{y}_0))$$

arbitrarily large.

Further, as noted earlier, $B(e)$ and $B(e_1)$ are independent of $\boldsymbol{y}$.

Therefore, at $\lambda \boldsymbol{y}_0$, for λ suitably large, $e(s,\hat{\boldsymbol{y}})$ can be made to perform better than $e_1(s,\hat{\boldsymbol{y}})$.

In other words,

$$V(e(s,\hat{\boldsymbol{y}}),\lambda \boldsymbol{y}_0) < V(e_1(s,\hat{\boldsymbol{y}}),\lambda \boldsymbol{y}_0),$$

which is a contradiction. Hence, the result. □

We now move on to prove a result pertaining to the nonexistence of uniformly minimum variance linear unbiased estimator (UMVLUE).

Let p be a given sampling design with

$$\pi_i = \sum_{s \ni i} p(s) > 0 \quad \forall i = 1,2,\ldots,N.$$

Based on p the Horvitz–Thompson estimator for $T(\boldsymbol{y})$ is given by

$$e_{\mathrm{HT}}(s,\boldsymbol{y}) = \sum_{i \in s} \frac{y_i}{\pi_i}.$$

Theorem 4.2. *There does not exist a UMVLU estimator for the total* $T(\boldsymbol{y}) = \sum_{i=1}^{N} y_i$ *in the RR set-up.*

Proof. A linear estimator is of the type $e(s,\boldsymbol{y}) = \sum_{i=1}^{N} b(s,i) y_i$.

We first show that for a given design p the term $B(e)$ of (4.1) is minimized for the Horvitz–Thompson estimator in the class of all linear unbiased estimators.

By Cauchy–Schwarz inequality

$$\sum_{s \ni i} b^2(s,i)p(s) \sum_{s \ni i} p(s) \geqslant \left\{ \sum_{s \ni i} b(s,i)p(s) \right\}^2 \quad \forall i = 1,2,\ldots,N$$

or equivalently invoking the condition of unbiasedness we have

$$\sum_{s \ni i} b^2(s,i)p(s) \geqslant \frac{1}{\pi_i}, \quad 1 \leqslant i \leqslant N,$$

and the equality is attained if and only if

$$b(s,i) = \frac{1}{\pi_i} \quad \forall s \ni i,\ 1 \leqslant i \leqslant N.$$

Thus, the Horvitz–Thompson estimator minimizes $B(e) = \sum_{i=1}^{N} E_p(b^2(s,i))$.

If possible let there exist an estimator e_0 that is UMVLUE.

(a) $e_0 \neq e_{\mathrm{HT}}$: Since $B(e_{\mathrm{HT}}) < B(e_0)$, in view of (4.2), we must have

$$V_p(e_{\mathrm{HT}}(s,\boldsymbol{y})) > V_p(e_0(s,\boldsymbol{y})) \quad \forall \boldsymbol{y}.$$

But this would mean that e_{HT} is inadmissible in the direct response set-up, which is a contradiction.

(b) $e_0 = e_{HT}$: Consider any other admissible estimator e_1 in the direct response set-up. This estimator would be better than e_{HT} at some vector, say $\boldsymbol{y_0}$, i.e.,

$$V_p(e_{HT}(s, \boldsymbol{y_0})) > V_p(e_1(s, \boldsymbol{y_0})).$$

Thus, choosing λ suitably we can make

$$V_p(e_{HT}(s, \lambda \boldsymbol{y_0})) - V_p(e_1(s, \lambda \boldsymbol{y_0}))$$

arbitrarily large.

Further $B(e_{HT})$ and $B(e_1)$ are free of $\boldsymbol{y}$. Hence at $\lambda \boldsymbol{y_0}$, for λ suitably large, $e_1(s, \hat{\boldsymbol{y}})$ can be made to perform better than $e_{HT}(s, \hat{\boldsymbol{y}})$.

In other words,

$$V(e_1(s, \hat{\boldsymbol{y}}), \lambda \boldsymbol{y}_0) < V(e_{HT}(s, \hat{\boldsymbol{y}}), \lambda \boldsymbol{y}_0),$$

which again is a contradiction.

Hence, UMVLUE does not exist. □

5. Superpopulation model

We now introduce the notion of superpopulation model.

Let $y_1, y_2, \ldots, y_N$ be a realization of the random variables $Y_1, Y_2, \ldots, Y_N$ with joint distribution specified by the first- and second-order moments as

$$E_\xi(Y_i) = \alpha_i, \quad 1 \leqslant i \leqslant N,$$

$$V_\xi(Y_i) = \tau_i^2, \quad 1 \leqslant i \leqslant N$$

and

$$\mathrm{Cov}_\xi(Y_i, Y_j) = 0, \quad 1 \leqslant i \neq j \leqslant N, \tag{5.1}$$

where $\tau_i^2 > 0$ and $\alpha_i, 1 \leqslant i \leqslant N$, are unknown parameters of model (5.1).

We know that

$$E_\xi V_p(e(s, \boldsymbol{y})) = V_p(e(s, \boldsymbol{y}), \boldsymbol{\alpha}) + \sum_{i=1}^{N} \tau_i^2 (E_p b^2(s, i) - 1).$$

Hence,

$$E_\xi V(e(s, \hat{\boldsymbol{y}})) = V_p(e(s, \boldsymbol{y}), \boldsymbol{\alpha}) + \sum_{i=1}^{N} (\tau_i^2 + \sigma^2) E_p(b^2(s, i)) - \sum_{i=1}^{N} \tau_i^2.$$

As the above expression is infested with too many unknown parameters there is no possibility of optimizing it. We can, however, state the following theoretical result.

Theorem 5.1. *If $(p, e^*(s, \boldsymbol{y}))$ is the best linear unbiased strategy, in the direct response set-up, in the sense of minimum $E_\xi V_p(e(s, \boldsymbol{y}))$ then $(p, e^*(s, \hat{\boldsymbol{y}}))$ would be better than $(p_1, e_1(s, \hat{\boldsymbol{y}}))$ in the RR set-up, in the sense of smaller $E_\xi V(e(s, \hat{\boldsymbol{y}}))$ if*

$$E_\xi(V_p(p_1, e_1) - V_p(p, e^*)) \geqslant \sigma^2 \{B(p, e^*) - B(p_1, e_1)\}.$$

Remark 5.1. It is well known that in certain situations under appropriate super-population models optimal strategies to estimate $T(\boldsymbol{y})$ exist in the direct response set-up. Some of these results smoothly carry over to the traditional RR set-up (see, e.g., Adhikary et al., 1984; Chaudhuri, 1987). The proposed RR set-up, however, does not admit such a carry over.

6. Comparing the two types of randomization

We finally move on to the comparison of two types of randomizations.

To make meaningful comparisons we assume that in the traditional RR set-up the ith individual either reports the true value y_i or generates and reports an observation on the random variable X, that has the same support as that of y, with specified probabilities. To make things formal let us first define

$$Z_i = \begin{cases} y_i \text{ with prob } \theta \\ X_i \text{ with prob } 1-\theta, \end{cases}$$

where $0 < \theta \leqslant 1$ and $X_1, X_2, \ldots, X_N$ are i.i.d. with d.f. F, mean μ and variance σ^2. We assume that the support of the random variable X is the same as that of y.

Note that $\theta = 1$, in the above framework, corresponds to direct response set-up. To compute the moments of Z_i, as mentioned earlier there are two stages of randomizations, at the first stage it is decided whether to use the variable y or X, if it is X, at the second stage we draw an observation from the distribution of X.

It is, therefore, easy to see that Z_i has expectation

$$\theta y_i + (1-\theta)\mu$$

and variance

$$\theta(1-\theta)(y_i-\mu)^2 + (1-\theta)\sigma^2 = \theta^2\sigma^2 w_i \quad \text{(say)},$$

where

$$w_i = \frac{1-\theta}{\theta}\left(\frac{y_i-\mu}{\sigma}\right)^2 + \frac{1-\theta}{\theta^2}. \tag{6.1}$$

For $\theta = 1$, e.g., $w_i = 0$, $\forall i = 1, 2, \ldots, N$ and that corresponds to direct response set-up.

Observe that y_i can be estimated unbiasedly by

$$\hat{\hat{y}}_i = \frac{Z_i - (1-\theta)\mu}{\theta}. \tag{6.2}$$

Hence, the estimator

$$e(s, \hat{\hat{\boldsymbol{y}}}) = \sum_{i=1}^{N} b(s,i)\hat{\hat{y}}_i \tag{6.3}$$

would be unbiased for $T(\boldsymbol{y})$.

We again use the symbolic formula $V = V_1E_2 + E_1V_2$ to compute the variability of estimator (6.3),

$$V(e(s,\hat{\boldsymbol{y}})) = V_p(e(s,\boldsymbol{y})) + E_pV_R\left(\sum_{i=1}^{N} b(s,i)\hat{y}_i\right)$$

$$= V_p(e(s,\boldsymbol{y})) + E_p\left(\sum_{i=1}^{N} b^2(s,i)V_R\left(\frac{Z_i}{\theta}\right)\right).$$

Thus,

$$V(e(s,\hat{\boldsymbol{y}})) = V_p(e(s,\boldsymbol{y})) + \sigma^2\sum_{i=1}^{N} w_iE_p\left(b^2(s,i)\right), \tag{6.4}$$

where w_i's are given by (6.1).

To compare the proposed method of randomization with the traditional one we compare the variabilities of estimators (3.2) and (6.3), i.e., we compare expressions (3.3) and (6.4). This, in turn, reduces to comparing the expressions

$$\sigma^2\sum_{i=1}^{N} E_p(b^2(s,i)) \quad \text{and} \quad \sigma^2\sum_{i=1}^{N} w_iE_p(b^2(s,i)),$$

where again w_i's are given by (6.1).

We now have the following theorem.

Theorem 6.1. *If*

$$\theta \leqslant \theta_0 = \frac{\sqrt{5}-1}{2} \approx 0.618, \tag{6.5}$$

then the new randomization method is better than the traditional one.

Proof. $\theta \leqslant \theta_0 \Rightarrow 1 \leqslant (1-\theta)/\theta^2$. Hence, $\sigma^2 \leqslant [(1-\theta)/\theta^2]\sigma^2 \leqslant \sigma^2 w_i \ \forall i$. Therefore, $\sigma^2\sum_{i=1}^{N} E_p(b^2(s,i)) \leqslant \sigma^2\sum_{i=1}^{N} w_iE_p(b^2(s,i))$. Hence the result. □

Remark 6.1. Though $\theta_0 \approx 0.618$, (6.5) is only a sufficient condition and it does not take into account the term $[(1-\theta)/\theta]((y_i-\mu)/\sigma)^2$ that could be quite arbitrary. Thus, since w_i's depend on y_i's the new method is likely to be better than the traditional one even when θ is actually somewhat larger than θ_0.

Remark 6.2. There are natural restrictions on the choice of X in the traditional RR set-up. For example, if the study variable y is nonnegative then X too has to be nonnegative or if the study variable y is a binary variable then X too has to be a binary variable. This is necessary so that the Statistician is unable to know whether the reported value is respondent's y-value or X-value. In the proposed RR set-up, however, there is no need for such restrictions.

We, finally, have the following theorem.

Theorem 6.2. *Consider the traditional RR set-up that uses random variable X_1 with mean μ_1 and variance σ_1^2. If we choose a random variable X with mean μ and variance σ^2 in the new RR set-up such that*

$$\sigma^2 \leqslant \frac{1-\theta}{\theta^2}\sigma_1^2,$$

then the new RR set-up is better than the traditional one.

Proof. The proof is similar to that of Theorem 6.1. □

In view of Remark 6.1 and as w_i's involve y_i's a similar statement cannot be made in the reverse direction.

Remark 6.3. The findings of this section can easily be adapted to compare any estimator belonging to the class of estimators (3.13) in the traditional RR set-up with its counterpart in the proposed RR set-up.

Acknowledgements

The authors thank the editors for their comments that led to improvement over the earlier version of the paper.

References

Adhikary, A.K., Chaudhuri, A., Vijayan, K., 1984. Optimum sampling strategies for randomized response trials. Internat. Statist. Rev. 52, 115–125.

Chaudhuri, A., 1987. Randomized response surveys of finite populations: A unified approach with quantitative data. J. Statist. Plann. Inference 15, 157–165.

Chaudhuri, A., Mukerjee, R., 1987. Randomized Response – Theory and Techniques. Marcel Dekker, New York.

Hedayat, A.S., Sinha, B.K., 1991. Design and Inference in Finite Population Sampling. Wiley, New York.