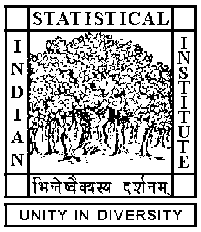
Enhancing a Randomized Response Model to Estimate the Sensitive Quantitative Population Mean

Technical Report No: ISI/ASD/2011/7
Dated: 28 December 2011

Kajal Dihidar
and
Joydeep Chowdhury

Applied Statistics Unit
Indian Statistical Institute
Kolkata 700108, India

Indian Statistical Institute
Applied Statistics Unit
Kolkata 700108, India



Technical Report No: ISI/ASD/2011/7

Enhancing a randomized response model to estimate the sensitive quantitative population mean

Kajal Dihidar[1] and Joydeep Chowdhury[2]
Indian Statistical Institute
Kolkata, West Bengal, India

Abstract

In this paper, we consider estimating population mean of sensitive quantitative variables, like expenditure on alcohol, number of female feticides, amount of dowry, etc. Gjestvang and Singh (2009) proposed a randomized response model to estimate sensitive quantitative population mean based on selecting the respondents by simple random sampling with replacement scheme. The specific advantage of this model over earlier models like Himmelfarb and Edgell (1980), Eichorn and Hayre (1983), and others, is that unlike usual models, Gjestvang and Singh (2009)'s model has estimator and variance estimator free from the known parameters of scrambling variable. In our present work, we have modified Gjestvang and Singh (2009)'s randomized response technique and have shown that, whatever be the sampling design, our modified estimator performs better than the usual estimator. Considering general varying probability sampling design, we have obtained unbiased estimator of population mean and unbiased variance estimator. Finally, we present a numerical exercise.

AMS Subject Classification: 62 D05.

Key words and phrases: Efficiency in estimation; Sensitive attribute; Unequal probability sampling; Scrambling variable.

1 Introduction

A challenging problem faced in conducting surveys is to gather reliable data on stigmatizing or sensitive topics such as drug addiction, induced abortion, drunken driving, habitual tax evasion, excessive gambling, etc. This is difficult since many respondents may be reluctant to truthfully answer direct questions on such topics. To overcome this difficulty, Warner (1965) pioneered a randomized response technique for estimating the proportion of persons having a particular sensitive characteristic in a community, on the basis of a simple random sampling with replacement scheme of respondents. After that, in the last 50 years, the literature on randomized responses has become rich. Different techniques dealing with different situations, extensions to quantitative data, unequal probability sampling even without replacement have been spawned.

Himmelfarb and Edgell (1980) considered the additive randomized response technique to estimate the population mean of a sensitive quantitative variable in a community, for example, to estimate the average monthly expenditure due to alcohol per household, or to estimate the average number of female feticide occurred per household within a particular time period, or to estimate the average amount of money taken as dowry per household, etc.

[1]Corresponding author. Email id: dkajal@isical.ac.in, kajaldihidar@gmail.com
[2]Email id: joydeepchowdhury01@gmail.com

Their procedure involves the reporting of a randomized response by adding the actual value with a random number drawn from a known distribution of a scrambling variable having known mean and variance. Later Eichorn and Hayre (1983) developed multiplicative model based on the reporting of a randomized response by multiplying the actual value with a random number drawn from the known distribution of the scrambling variable. In both models, the estimators depend upon the parameters of the scrambling variable. Gjestvang and Singh (2009) proposed a randomized response model to estimate sensitive population mean based on SRSWR scheme of respondents where the estimators and the variance estimators are free from the parameters of the randomization device. In this paper, we extend this model to general sampling schemes beyond SRSWR because we believe that a large-scale survey involves many items including merely a few of which may be sensitive ones, a sample having been drawn judiciously for estimation covering each. No separate sample is needed to tackle sensitive issues.

Mangat and Singh (1990) gave a modification to Warner's (1965) model to estimate the sensitive population proportion in a community and studied conditions under which this modification led to improved efficiency in estimation. Traditionally, these models are based on SRSWR. In this paper, we show that whatever be the sampling design for selecting the respondents, Mangat and Singh (1990)'s modification can be profitably applied to Gjestvang and Singh (2009)'s model for estimating sensitive quantitative population mean.

Optional randomized response techniques are available in literature in two ways. The first type (Chaudhuri and Mukerjee, 1985, 1988; Chaudhuri and Saha, 2005) permits the intentional disclosure of truth overlooking the stigma. The second type (Mangat and Singh, 1994; Singh and Joarder, 1997; Gupta et al., 2002; Arnab, 2004; Pal, 2008; Chaudhuri and Dihidar, 2009) allows the respondent to either reveal the true characteristic or to follow a prescribed randomized response device unnoticed by the interviewer. Here we consider this second type of optional randomized response technique. In this paper, we show that irrespctive of how a sample of respondents is chosen the optional randomized response tecnique can be gainfully applied to Gjestvang and Singh (2009)'s model.

We organize our findings in the following sections. In Section 2, we present Mangat and Singh (1990)'s modification and optional randomized response technique applied to Gjestvang and Singh (2009)'s model. In Section 3, we derive unbiased estimators and variance estimators based on the samples chosen by varying probabilities. Next we illustrate our findings through a numerical presentation showing the gains in efficiencies in Section 4. Finally in Section 5, we give the concluding remaks.

2 Gjestvang and Singh (2009)'s model and its modification

Suppose in a finite survey population $U = (1, \ldots, i, \ldots, N)$ a person labelled i has the value y_i defined on a stigmatizing quantitative variable Y. Our problem is to estimate $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

The Gjestvang and Singh (2009)'s technique of reporting randomized response has two steps. First let Z be a scrambling variable of known distribution having known mean and variance as μ_z and σ_z^2. Each respondent selected by SRSWR scheme is requested to generate a value using a randomization device generating values from the distribution of Z. At the second step, for two known positive real numbers α and β, a randomization device, say a

deck of cards, is supplied where $\frac{\beta}{\alpha+\beta}$ proportion of the cards bear the statement 'Add the scrambling variable Z multiplied by α to your real value of Y' and the rest $\frac{\alpha}{\alpha+\beta}$ proportion of the cards bear the statement 'Subtract the scrambling variable Z multiplied by β to your real value of Y'. Each respondent is requested to draw one card secretly and report the scrambled response accordingly. So, for the person i, if z be the value generated from the known distribution of the variable Z, x_i be the randomized response reported by him then

$$\begin{aligned} x_i &= y_i + \alpha z \quad \text{with probability} \quad \tfrac{\beta}{\alpha+\beta}, \text{ and} \\ &= y_i - \beta z \quad \text{with probability} \quad \tfrac{\alpha}{\alpha+\beta}. \end{aligned}$$

Writing E_R, V_R and C_R to denote expectation, variance and covariance operators with respect to this randomization with $C_R(x_k, x_k') = 0, \ \ \forall \ \ k, k' \ \ (k \neq k')$, one has

$$E_R(x_i) = \frac{\beta}{\alpha+\beta}(y_i + \alpha\mu_z) + \frac{\alpha}{\alpha+\beta}(y_i - \beta\mu_z) = y_i, \tag{2.1}$$

and

$$V_R(x_i) = E_R(x_i^2) - (E_R(x_i))^2$$

$$= \frac{\beta}{\alpha+\beta}\left[y_i^2 + \alpha^2\sigma_z^2 + \alpha^2\mu_z^2 + 2y_i\alpha\mu_z\right] + \frac{\alpha}{\alpha+\beta}\left[y_i^2 + \alpha^2\sigma_z^2 + \alpha^2\mu_z^2 - 2y_i\alpha\mu_z\right] - y_i^2$$

$$= y_i^2 + \alpha\beta\left(\sigma_z^2 + \mu_z^2\right) - y_i^2 = \alpha\beta\left(\sigma_z^2 + \mu_z^2\right) = \Phi_{GS}, \text{ say .} \tag{2.2}$$

Let E_P, V_P denote expectation and variance operators with respect to the sampling design for selection of respondents. Now if x_k denotes the randomized response reported by the respondent chosen at the k_{th} draw, then Gjestvang and Singh (2009)'s unbiased estimator for $\bar{Y}$ based on SRSWR of n draws is

$$\hat{\bar{Y}}_{GS} = \frac{1}{n}\sum_{k=1}^{n} x_k, \tag{2.3}$$

since

$$E(\hat{\bar{Y}}_{GS}) = E_P E_R(\hat{\bar{Y}}_{GS}) = E_P(\frac{1}{n}\sum_{k=1}^{n} y_k) = \bar{Y}.$$

$$V(\hat{\bar{Y}}_{GS}) = V_P E_R(\hat{\bar{Y}}_{GS}) + E_P V_R(\hat{\bar{Y}}_{GS}) = V_P(\frac{1}{n}\sum_{i=1}^{n} y_k) + E_P(\frac{\Phi_{GS}}{n}) = \frac{\sigma^2}{n} + \frac{\Phi_{GS}}{n}, \tag{2.4}$$

where $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2$. The unbiased variance estimator for $\hat{\bar{Y}}_{GS}$ is

$$v(\hat{\bar{Y}}_{GS}) = \frac{1}{n(n-1)}\sum_{k=1}^{n}(x_k - \hat{\bar{Y}}_{GS})^2, \tag{2.5}$$

since

$$E_R(v(\hat{\bar{Y}}_{GS})) = \frac{1}{n(n-1)} E_R\left[\sum_{k=1}^{n} x_k^2 - n\hat{\bar{Y}}_{GS}^2\right]$$

$$= \frac{1}{n(n-1)} \left[\sum_{k=1}^{n} \left(V_R(x_k) + E_R^2(x_k) \right) - n \left(V_R(\hat{\bar{Y}}_{GS}) + E_R^2(\hat{\bar{Y}}_{GS}) \right) \right]$$

$$= \frac{1}{n(n-1)} \left[n\Phi_{GS} + \sum_{k=1}^{n} y_k^2 - n\frac{\Phi_{GS}}{n} - n \left(\frac{1}{n} \sum_{k=1}^{n} y_k \right)^2 \right]$$

$$= \frac{\Phi_{GS}}{n} + \frac{1}{n} \frac{\sum_{k=1}^{n} \left(y_k - \frac{1}{n} \sum_{k=1}^{n} y_k \right)^2}{(n-1)}.$$

Then under SRSWR,

$$E_P E_R(v(\hat{\bar{Y}}_{GS})) = \frac{\Phi_{GS}}{n} + \frac{\sigma^2}{n} = V(\hat{\bar{Y}}_{GS}).$$

The specific advantage of this model over the traditional additive and multiplicative models is that, for this model the estimator for the population mean and the variance estimator for that are free from the parameters of the randomization device and hence can be safely used once the choice of two known real valued constants α and β is decided, whereas for the traditional models the estimator depends upon the parameters of the scrambling variables.

2.1 Mangat and Singh (1990)'s modification applied to Gjestvang and Singh (2009)'s model

Mangat and Singh (1990) modified Warner (1965)'s model by giving a chance to the selected respondents to reveal their true values by a lottery method. They studied the condition when their modified estimator performs better than Warner (1990)'s usual estimator. Motivated by their approach, we modify Gjestvang and Singh (2009)'s model in the following way.

A box namely Tbox is proposed to be made consisting of two types of cards: a proportion T of cards being marked 'True' and the remaining cards marked 'RR'. A sampled person i is asked to first draw a card from Tbox. He/she is requested to report the true value y_i if a 'True' marked card is drawn, and otherwise to use Gjestvang and Singh (2009)'s randomized response device and to produce a response x_i. The entire process will be unnoticed by the interviewer, thus maintaining the respondent's protection of privacy. Hence, the response from the i_{th} person sampled is

$$\begin{aligned} z_i &= y_i \quad \text{if a 'True' card is drawn from Tbox,} \\ &= x_i \quad \text{if an 'RR' card is drawn instead.} \end{aligned}$$

Clearly following Eq.(2.1),

$$E_R(z_i) = Ty_i + (1-T)y_i = y_i. \tag{2.6}$$

And

$$V_R(z_i) = E_R(z_i^2) - (E_R(z_i))^2$$

$$= Ty_i^2 + (1-T)E_R(x_i^2) - y_i^2$$

$$= (1-T)[V_R(x_i) + (E_R(x_i))^2] - (1-T)y_i^2$$

$$= (1-T)[\Phi_{GS} + y_i^2] - (1-T)y_i^2 = (1-T)\Phi_{GS} = \quad \text{a known constant .} \tag{2.7}$$

In Section 3, we will show that when the sample s of respondents is chosen with probabilities $p(s)$, p being any arbitrary sampling design, and Gjestvang and Singh (2009)'s model is used to estimate the population mean, the estimator is a function of x_i values for the original model and the same function of z_i under its modified version following Mangat and Singh (1990). More importantly, the variance of each of these estimators will be a sum of two terms, the first term being a constant depending on the design p and the second part a function of the varianes of either x_i's or z_i's, as the case may be. So, in order to compare the variances of the estimators under the original model and its modified version, it is enough to compare the varances of x_i and z_i.

Now, since $0 < T < 1$, on comparing the variances of the responses x_i and z_i under Gjestvang and Singh (2009)'s model and its modification following Mangat and Singh (1990), we have

$$V_R(z_i) \leq V_R(x_i) \quad \forall \quad i \in U. \tag{2.8}$$

So, using the same scrambling variable, the modified model by giving a chance to the selected respondent to reveal his/her true value by a lottery method is always better than the original model given by Gjestvang and Singh (2009).

2.2 Optional randomized response technique applied to Gjestvang and Singh (2009)'s model

Let us suppose, an unknowable probability $C_i (0 \leq C_i \leq 1)$ has been assigned by nature to the person i that he/she gives out the true value y_i if sampled and addressed. With probability $(1 - C_i)$ he/she is supposed to give out the randomized response following Gjestvang and Singh (2009)'s technique. Thus let his/her response reported to the interviewer be u_i, hiding the type of his response (direct or randomized) to the interviewer, and hence protecting his privacy. Additionally, he/she is requested to give the responses twice and independently, i.e. u_{1i} and u_{2i}, say. That means,

$$\begin{aligned} u_{1i} &= y_i \quad \text{with probability} \quad C_i \\ &= x_{1i} \quad \text{with probability} \quad (1 - C_i) \end{aligned}$$

and

$$\begin{aligned} u_{2i} &= y_i \quad \text{with probability} \quad C_i \\ &= x_{2i} \quad \text{with probability} \quad (1 - C_i) \end{aligned}$$

are two independent randomized responses obtained from person i.

Then clearly following Eq.(2.1),

$$E_R(u_{1i}) = C_i y_i + (1 - C_i) y_i = y_i, \quad \text{and} \quad E_R(u_{2i}) = C_i y_i + (1 - C_i) y_i = y_i, \tag{2.9}$$

and proceeding as for Eq. (2.7),

$$V_R(u_{1i}) = (1 - C_i)\Phi_{GS} = V_R(u_{2i}). \tag{2.10}$$

We note that $V_R(u_{1i})$ and $V_R(u_{2i})$ are the same unknown constants as C_i is unknown to us. So, if

$$u_i = \frac{u_{1i} + u_{2i}}{2}, \quad \text{then } E_R(u_i) = y_i, \quad \text{and} \quad V_R(u_i) = \frac{(1 - C_i)\Phi_{GS}}{2}, \tag{2.11}$$

and for an unbiased estimator of $V_R(u_i)$, we may take $v_i = \dfrac{(u_{1i} - u_{2i})^2}{4}$ as $E_R(v_i) = V_R(u_i)$.

Now, since $0 \leq C_i \leq 1$, on comparing the variances of the responses x_i and u_i under Gjestvang and Singh (2009)'s model and its optional randomized response modification following, we have

$$V_R(u_i) \leq V_R(x_i) \quad \forall \quad i \in U. \tag{2.12}$$

So, using the same scrambling variable, the modified model by letting free the selected respondent to reveal his/her true value depending on his/her own desire is always better than the original model given by Gjestvang and Singh (2009).

It is important to note that in dealing with optional randomized response modification to Gjestvang and Singh (2009)'s model two independent responses are needed while for Mangat and Singh (1990)'s modification only one response suffices for each sampled person.

3 Unbiased estimators and variance estimators based on RR's obtained from samples chosen by varying probabilities

With y_i as defined in the preceding sections, we consider the estimation of the population mean $\bar{Y} = \sum_{i=1}^{N} y_i/N = Y/N$ for a quantitative stigmatizing variable, where Y is the population total of interest. As N is known to us, it reduces to the problem of estimating Y and we consider this problem based on a sample s chosen from U with probabilities $p(s)$ by any suitable sampling design p.

If direct responses y_i's are available for the respondents in s, one may use a homogeneous linear unbiased estimator of the form

$$\hat{Y}_b = \sum_{i \in s} y_i b_{si}, \tag{3.1}$$

where b_{si}'s are free of y_i's and satisfy $\sum_{s \ni i} p(s) b_{si} = 1, \forall\ i \in U$. Then it is well known that

$$E_P(\hat{Y}_b) = Y \text{ and } V_P(\hat{Y}_b) = \sum_{i=1}^{N} y_i^2 c_i + \sum \sum_{i,j=1, i \neq j}^{N} y_i y_j c_{ij} \tag{3.2}$$

where $c_i = E_P(b_{si}^2 I_{si}) - 1$ and $c_{ij} = E_P(b_{si} b_{sj} I_{sij}) - 1$ where I_{si} and I_{sij} are defined as

$$\begin{aligned} I_{si} &= 1 \text{ if } i \in s \\ &= 0, \text{ otherwise,} \end{aligned}$$

and $I_{sij} = I_{si} I_{sj}$. Let c_{si} and c_{sij} be such that $E_P(c_{si} I_{si}) = c_i$ and $E_P(c_{sij} I_{sij}) = c_{ij}$. Then it follows that

$$E_P \left[\sum_{i \in s} y_i^2 c_{si} + \sum \sum_{i \neq j, i, j \in s} y_i y_j c_{sij} \right] = V_P(\hat{Y}_b) = E_P[\hat{V}_P(\hat{Y}_b)], \tag{3.3}$$

say. Thus $\hat{V}_P(\hat{Y}_b) = \sum_{i \in s} y_i^2 c_{si} + \sum \sum_{i \neq j, i, j \in s} y_i y_j c_{sij}$ is an unbiased estimator for $V_P(\hat{Y}_b)$.

3.1 Unbiased estimators based on RR's and their variances

When instead of y_i's, only the RR's under Gjestvang and Singh (2009)'s model discussed in Section 2 are available, we use the RR's to form the following estimator for Y

$$e_{GS} = \sum_{i \in s} x_i b_{si}. \tag{3.4}$$

Since $E_R(x_i) = y_i$, from (3.1) and (3.2) it follows that $E_P E_R(e_{GS}) = Y$. Thus e_{GS} is unbiased for Y.

The estimator based on the RR's obtained from Mangat and Singh (1990)'s modification of the corresponding model will be

$$e_{GST} = \sum_{i \in s} z_i b_{si} \tag{3.5}$$

and as $E_R(z_i) = y_i$, e_{GST} is unbiased for Y.

The estimator based on the RR's obtained from optional randomized response technique of the corresponding model will be

$$e_{GSO} = \sum_{i \in s} u_i b_{si} \tag{3.6}$$

and again, as $E_R(u_i) = y_i$, e_{GSO} is unbiased for Y.

To obtain the variance of e_{GS} we note that

$$V(e_{GS}) = E[e_{GS} - Y]^2 = E_P E_R[e_{GS} - Y]^2 = V_P(\hat{Y}_b) + E_P(\sum_{i \in s} b_{si}^2 V_R(x_i)), \tag{3.7}$$

where the first term is as in (3.2) and depends on the design p, while the $V_R(x_i)$ in the second term is specific to the model used and is as given in (2.2). Similarly, for the modified version of the model,

$$V(e_{GST}) = V_P(\hat{Y}_b) + E_P(\sum_{i \in s} b_{si}^2 V_R(z_i)), \tag{3.8}$$

where the $V_R(z_i)$ for the modified version of the model is as given in (2.7). For optional randomized response technique of the model,

$$V(e_{GSO}) = V_P(\hat{Y}_b) + E_P(\sum_{i \in s} b_{si}^2 V_R(u_i)), \tag{3.9}$$

where the $V_R(u_i)$ for this modified version of the model is as given in (2.11).

From (3.7), (3.8) and (3.9) it is now evident that in order to study the relative performances of e_{GS} and e_{GST} and of e_{GS} and e_{GSO} by comparing their variances, it is enough to compare $V_R(x_i)$ with $V_R(z_i)$ and to compare $V_R(x_i)$ with $V_R(u_i)$, as was done in Section 2.

3.2 Unbiased variance estimators

We note that if direct responses y_i's were available, then from (3.3) we could use $\hat{V}_P(\hat{Y}_b)$. But since we have randomized responses x_i's or z_i's or u_i's instead of y_i's, then to estimate unbiasedly the variances of e_{GS}, e_{GST} and e_{GSO}, we follow Raj (1966). We present the

formulation only for e_{GS} using the randomized responses x_i's. Then just replacing the x_i's by z_i's and by u_i's, the unbiased variance estimators of e_{GST} and e_{GSO} will be obtained.

Using the form of $\hat{V}_P(\hat{Y}_b)$ in (3.3), let

$$v_1 = \sum_{i\in s} x_i^2 c_{si} + \sum \sum_{i\neq j, i,j\in s} x_i x_j c_{sij} + \sum_{i\in s} \hat{V}_R(x_i)(b_{si}^2 - c_{si}),$$

and

$$v_2 = \sum_{i\in s} x_i^2 c_{si} + \sum \sum_{i\neq j, i,j\in s} x_i x_j c_{sij} + \sum_{i\in s} \hat{V}_R(x_i) b_{si}. \tag{3.10}$$

Following Raj (1966), $E_P E_R(v_1) = V(e_{GS}) = E_P E_R(v_2)$. So both v_1 and v_2 are unbiased variance estimators for e_{GS}.

Now for usual Gjestvang and Singh (2009)'s model and its Mangat and Singh (1990)'s modification, it is clear from (2.2) and (2.7) that both $V_R(x_i)$ and $V_R(z_i)$ are constants depending on the parameters of the devices, namely α, β, μ_z, σ_z and T. So for this model, both the variances $V_R(x_i)$ and $V_R(z_i)$ are known. Hence from (3.10), it is clear that those known constants serve as the unbiased estimators for $V_R(x_i)$ and $V_R(z_i)$ to be utilized as parts for estimating $V(e_{GS})$, $V(e_{GST})$. On the contrary, for the optional randomization modification, $V_R(u_i)$ is an unknown constant. Hence from (3.10), it is clear that in order to estimate $V(e_{GSO})$, first of all, we need to estimate $V_R(u_i)$ for all $i \in s$ and based on two independent randomized responses, $\hat{V}_R(u_i) = \dfrac{(u_{1i} - u_{2i})^2}{4} = v_i$ serves as unbiased estimator for $V_R(u_i)$ for $i \in s$.

Hence on proper substitution of $\hat{V}_R(x_i)$, $\hat{V}_R(z_i)$, and $\hat{V}_R(u_i)$ along with the respective randomized responses in Eq. (3.10) we can estimate unbiasedly the variances of e_{GS}, e_{GST} and e_{GSO}.

4 Numerical illustrations showing the gains in efficiencies

We note from Section 2, that both the modifications namely Mangat and Singh (1990)'s and optional randomization modification, lead Gjestvang and Singh (2009)'s model to improved versions uniformly. Inspite of that, in order to demonstrate how our method can be effectively applied in a practical survey situation, we present a numerical study. We consider a fictitious population data that has been used in Chaudhuri and Dihidar (2009). This population data is about $N = 117$ people with (1) last month's household expenses (E) in an appropriate currency and (2) the person's last month's expenses on purchase of alcohol (y). These values are displayed in Table 1. In order to assume the probabilities that may be assigned by nature to all the persons that every person gives out his/her true value y_i if sampled and addressed, we draw independently for each of the $N = 117$ persons a random number in between $(0, 1)$ rounded upto 2 decimal places and call them as C_i, for $i = 1, \ldots, 117$.

Our objective is to estimate the actual mean expenditure $\bar{Y}$ due to alcohol based on a sample drawn by an unequal probability without replacement sampling scheme from that population.

Using E values as size measures for these N people and the corresponding normed size measures namely p_i's $(0 < p_i < 1, i = 1, \ldots, 117, \sum_{i=1}^{N} p_i = 1)$ we draw from them a sample of $n = 25$ people employing Rao, Hartley and Cochran's sampling scheme (RHC, 1962). The RHC-scheme is applied by forming n groups with N_i units in the ith group by simple random sampling without replacement (SRSWOR) out of the N units such that, writing Σ_n as sum over n groups, $\Sigma_n N_i = N$. Optimal group sizes as given by Rao, Hartley and Cochran (1962) are

$$\begin{aligned} N_i &= [\tfrac{N}{n}] \text{ for } i = 1, \ldots, k \\ &= [\tfrac{N}{n}] + 1 \text{ for } i = k+1, \ldots, n. \end{aligned}$$

k is to be determined by solving $\Sigma_n N_i = N$. Denoting by $(p_{i1}, p_{i2}, \ldots, p_{iN_i})$ the normed size measures of the N_i units falling in the ith group, by Q_i the sum of those normed size measures, a unit i_k, say, from the ith group is chosen with probability p_{i_k}/Q_i. This is independently repeated across the n groups. For simplicity, we call the value obtained from the selected unit of i_{th} group as y_i and the normed size measure corresponding to the selected unit from i_{th} group as p_i. Then RHC's unbiased estimator for $Y = \Sigma_{i=1}^{N} y_i$ is

$$t = \Sigma_n \frac{Q_i}{p_i} y_i \tag{4.1}$$

and

$$V(t) = \frac{\Sigma_n N_i^2 - N}{N(N-1)} \left[\sum_{i<}^{N} \sum_{j=i}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 \right]. \tag{4.2}$$

A uniformly non-negative unbiased estimator of $V(t)$ as given by Rao, Hartley and Cochran (1962) is

$$v_P(t) = \frac{\Sigma_n N_i^2 - N}{N^2 - \Sigma_n N_i^2} \left[\Sigma_n \Sigma_n Q_i Q_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 \right], \tag{4.3}$$

where $\Sigma_n \Sigma_n$ denotes the summation over non-overlapping pairs of n groups.

However, the y_i's are not directly ascertainable in randomized response surveys and our problem is to estimate $\bar{Y}$. So, under this RHC sampling scheme, unbiased estimators for $\bar{Y}$ for the classical Gjestvang and Singh (2009)'s models and also its modified versions are obtained from (3.4), (3.5) and (3.6) as

$$e_1 = \frac{1}{N} \left[\sum_n \frac{Q_i}{p_i} x_i \right], \quad e_2 = \frac{1}{N} \left[\sum_n \frac{Q_i}{p_i} z_i \right], \quad \text{and} \quad e_3 = \frac{1}{N} \left[\sum_n \frac{Q_i}{p_i} u_i \right] \tag{4.4}$$

respectively, with $b_{si} = Q_i/p_i$, and x_i, z_i and u_i corresponding to the model itself and its two modifications being considered.

Now in order to compare the efficiencies of these alternative estimators, from Eqs. (3.7), (3.8), and (3.9), we obtain the variances of these estmators in the following way.

$$V(e_1) = V_P E_R \left(\frac{1}{N} \left[\sum_n \frac{Q_i}{p_i} x_i \right] \right) + E_P V_R \left(\frac{1}{N} \left[\sum_n \frac{Q_i}{p_i} x_i \right] \right)$$

$$= \frac{1}{N^2} \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \left[\sum_{i<}^{N} \sum_{j=i}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 \right] + \frac{\phi_{GS}}{N^2} E_P \left[\sum_{i=1}^{n} \frac{Q_i^2}{p_i^2} \right] = A + B, \text{ say.} \tag{4.5}$$

To obtain B, while performing the expectation E_P, we first perform expectation conditional to a fixed group G_i, say, then we perform expectation over all the possible groups. So,

$$B = \frac{\phi_{GS}}{N^2} E_G \left[\sum_{i=1}^{n} E\left(\frac{Q_i^2}{p_i^2} | G_i \right) \right] = \frac{\phi_{GS}}{N^2} E_G \left[\sum_{i=1}^{n} \sum_{j \in G_i} \frac{Q_j}{p_j} \right]$$

$$= \frac{\phi_{GS}}{N^2} E_G \left[\sum_{i=1}^{n} \sum_{j \in G_i} \frac{p_j + \sum_{k \neq j, k \in G_i} p_k}{p_j} \right]$$

$$= \frac{\phi_{GS}}{N^2} E_G \left[\sum_{i=1}^{n} \sum_{j \in G_i} 1 \right] + \frac{\phi_{GS}}{N^2} E_G \left[\sum_{i=1}^{n} \sum_{j \in G_i} \sum_{k \neq j, k \in G_i} \frac{p_k}{p_j} \right]$$

$$= \frac{\phi_{GS}}{N^2} N + \frac{\phi_{GS}}{N^2} \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \left[\sum_{j=1}^{N} \sum_{k \neq j}^{N} \frac{p_k}{p_j} \right]$$

$$= \frac{\phi_{GS}}{N} + \frac{\phi_{GS}}{N^2} \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \left[\sum_{j=1}^{N} \frac{1 - p_j}{p_j} \right]. \tag{4.6}$$

Thus from Eqs. (4.5) and (4.6) we have,

$$V(e_1) = \frac{1}{N^2} \left[\frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \sum_{i<}^{N} \sum_{j=i}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 + N \phi_{GS} + \phi_{GS} \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \sum_{j=1}^{N} \frac{1 - p_j}{p_j} \right]. \tag{4.7}$$

Similarly, as $E_R(z_i) = y_i$ and $V_R(z_i) = (1-T)\phi_{GS} = \phi_{GST}$, say, a known constant for all $i \in s$, proceeding in the same way, we have

$$V(e_2) = \frac{1}{N^2} \left[\frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \sum_{i<}^{N} \sum_{j=i}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 + N \phi_{GST} + \phi_{GST} \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \sum_{j=1}^{N} \frac{1 - p_j}{p_j} \right]. \tag{4.8}$$

For $V(e_3)$, we note that $E_R(u_i) = y_i$, but $V_R(u_i) = \dfrac{(1-C_i)\phi_{GS}}{2}$ is not constant and unknown for all $i \in s$. Hence, the first term of $V(e_3)$ is exactly equal to A and the second term B alalogous to Eq. (4.5) of $V(e_3)$ is given by:

$$\frac{\phi_{GS}}{2N^2} E_P \left[\sum_{i=1}^{n} \frac{Q_i^2}{p_i^2} (1 - C_i) \right] = B_{opt}, \text{ say.}$$

On completing the expectation, we obtain

$$B_{opt} = \frac{\phi_{GS}}{2N^2} \left(1 - \sum_{i=1}^{N} C_i \right) + \frac{\phi_{GS}}{2N^2} \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \left[\sum_{j=1}^{N} (1 - C_i) \frac{1 - p_j}{p_j} \right]. \tag{4.9}$$

Hence,

$$V(e_3) =$$

$$\frac{1}{N^2} \left[\frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \sum_{i<}^{N} \sum_{j=i}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 + \frac{\phi_{GS}}{2} (N - \sum_{i=1}^{N} C_i) + \frac{\phi_{GS}}{2} \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \sum_{j=1}^{N} (1 - C_j) \frac{1 - p_j}{p_j} \right]. \tag{4.10}$$

Now to measure the efficiency of e_2 and e_3 with respect to e_1, we compute, $Eff_{21} = \left(\frac{V(e_1)}{V(e_2)}\right) \times 100$ and $Eff_{31} = \left(\frac{V(e_1)}{V(e_3)}\right) \times 100$ for several illustrative values of the device parameters, namely of μ_z, σ_z, α, β, and T. We present the results in Table 2.

Next, to get some ideas about the estimates along with the measure of errors obtained in a practical sample survey situation, we perform the simulation by drawing samples of size $n = 25$ from the same population by RHC scheme with E values as the size measures. For each selected respondent, we generate the randomized responses for some illustrative values of the model parameters μ_z, σ_z, α, β, and T and we compute the three estimators e_1, e_2 and e_3. Following Eqs. (3.10), (4.3) and a result in Chaudhuri, Adhikary and Dihidar (2000), an unbiased variance estimator for e_1 is obtained as

$$\hat{V}(e_1) = v(e_1) = \frac{1}{N^2}\left[\frac{\sum_n N_i^2 - N}{N^2 - \sum_n N_i^2}\Sigma_n\Sigma_n Q_i Q_j \left(\frac{x_i}{p_i} - \frac{x_j}{p_j}\right)^2 + \sum_n \hat{V}_R(x_i)\frac{Q_i}{p_i}\right], \qquad (4.11)$$

where the x_i and $\hat{V}_R(x_i)$ correspond to the model being considered. By replacing x_i respectively by z_i and u_i and $\hat{V}_R(x_i)$ respectively by $\hat{V}_R(z_i)$ and $\hat{V}_R(u_i)$ in $v(e_1)$ above, we obtain unbiased variance estimator for e_2 and of e_3, namely $\hat{V}(e_2)$ and $\hat{V}(e_3)$.

We use the coefficient of variation (CV) as the criteria for our efficiency comparisons as it is well known that less the CV, the more efficient is the corresponding estimator. The coefficient of variation for a particular estimator is estimated as $\text{cv} = 100 \times \frac{\sqrt{\text{estimated variance}}}{\text{estimate}}$. In Table 3, for some chosen values of the device parameters, we present the performance of the estimator e_1 under the classical Gjestvang and Singh (2009)'s model and also the estimators e_2 and e_3 under its two modified versions as described in Section 2.

Table 1: A fictitious population of 117 persons

Person's sl.no. i	E_i	y_i	Person's sl.no. i	E_i	y_i
1	2891.31	492.31	61	2636.53	452.58
2	4261.13	722.69	62	1344.76	232.38
3	2262.45	0	63	1544.81	0
4	2530.49	424.09	64	1255.77	0
5	2430.49	413.75	65	1328.88	228.24
6	4226.83	722.85	66	3258.28	560.34
7	3270.41	559.66	67	2740.52	464.74
8	1179.95	204.70	68	4298.5	732.49
9	1902.73	0	69	2185.70	377.70
10	1482.09	0	70	251.27	42.54
11	1480.36	250.44	71	3065.67	523.67
12	250.9	47.80	72	1194.98	0
13	2255.33	0	73	179.98	0
14	2525.85	424.70	74	3845.06	651.45
15	1241.19	215.12	75	1188.66	0
16	1256.66	0	76	189.36	25.82
17	2194.89	374.59	77	1247.3	0
18	3187.48	540.80	78	5004.93	855.39
19	193.65	33.38	79	1505.03	249.29
20	1669.54	285.67	80	3240.26	554.56
21	3074.11	523.67	81	3254.33	548.27
22	4187.81	700.05	82	334.97	56.20
23	1264.92	227.93	83	1242.27	208.06
24	3196.59	541.03	84	4181.9	0
25	3354.57	568.83	85	187.78	30.37
26	2717.12	459.06	86	3242.91	543.94
27	2927.63	500.67	87	4334.62	734.94
28	4147.14	700.42	88	2575.97	436.85
29	3385.06	571.10	89	2608.09	446.20
30	2644.63	0	90	4703.93	809.93
31	2495.64	0	91	1940.05	337.61
32	4400.64	756.44	92	2724.16	459.22
33	3284.96	562.61	93	3199.71	536.66
34	1334.98	226.23	94	1241.56	203.21
35	1408.34	235.96	95	1173.01	192.27
36	1241.83	208.51	96	1435.06	247.81
37	4649.75	790.65	97	251.42	0
38	2243.53	374.68	98	3236.45	548.97
39	1120.97	184.68	99	1309.49	225.07
40	1296.67	220.31	100	3247.36	0

Table 1 continued:

Person's sl.no. i	E_i	y_i	Person's sl.no. i	E_i	y_i
41	2878	0	101	1271.32	209.80
42	1268.51	0	102	208.24	27.95
43	1258.95	212.02	103	246.96	39.35
44	2990.47	506.01	104	1474.4	255.89
45	1299.93	222.41	105	2430.23	417.67
46	205.55	35.79	106	1148.49	191.86
47	1245.97	216.11	107	640.08	101.62
48	1241.24	209.14	108	3942.96	670.74
49	195.59	34.77	109	2202.25	0
50	2260.59	379.27	110	241.63	31.99
51	242.99	39.17	111	4191.92	706.51
52	195.08	36.56	112	4269.03	726.91
53	3194.31	542.93	113	2742.73	466.04
54	2307.38	0	114	542.3	0
55	4842.01	823.37	115	1546.3	254.72
56	2904.35	0	116	1478.00	260.68
57	3154.77	544.98	117	789.00	134.62
58	2191.78	372.80			
59	2241.53	375.12			
60	1241.82	0			

For this population, $\bar{Y} = \dfrac{1}{117}\sum_{i=1}^{117} y_i = 304.47.$

Table 2. Efficiency comparisons of the modified Gjestvang and Singh (2009)'s model with respect to the classical model under RHC (1962) scheme

μ_z	σ_z	α	β	Eff_{21}									Eff_{31}
				T									
				0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
17	11	23	77	110.99	124.68	142.24	165.55	197.99	246.26	325.64	480.53	916.46	310.64
19	27	23	128	111.08	124.93	142.72	166.41	199.54	249.14	331.55	495.42	979.63	315.90
28	19	37	58	111.07	124.91	142.67	166.33	199.40	248.87	331.01	494.04	973.58	315.42

Table 3. Simulation results for classical Gjestvang and Singh (2009)'s model and its two modifications under RHC (1962) scheme

μ_z	σ_z	α	β	cv(%) of e_1	cv(%) of e_2									cv(%) of e_3
					T									
					0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
17	11	23	77	32.03	17.55	17.36	16.27	19.38	15.32	14.82	11.29	10.25	8.28	19.55
19	27	23	128	37.85	28.96	27.89	24.11	20.85	21.33	19.84	22.30	20.09	17.18	15.72
28	19	37	58	48.00	29.99	31.86	29.92	36.44	29.15	30.57	31.73	24.95	18.59	19.63

5 Concluding remarks

We observe from the efficiency comparisons in Table 2 that Gjestvang and Singh (2009)'s model considered in this paper can be gainfully modified with a considerable amount of gain by allowing the respondents to reveal their true values in both ways, one way by lottery method and the other way by depending on their own desires. Moreover, the estimated coefficient of variations obtained from simulation results as depicted in Table 3, are showing smaller values for both types of modifications in comparison to the classical method, thus proving the usefulness of the modifications in practical sample survey situations. So, based on the theoretical results and on the numerical results of Table 2 and 3, we can conclude that one should not hesitate to give a chance to the respondent to reveal their true values. This chance may be determined by lottery method or may be by letting the respondents free to reveal his/her true value depending on his/her own desire, unnoticed by the interviewer. Both methods are proved to have assured gain in efficiencies.

References

Arnab, Raghunath (2004). Optional randomized response techniques for complex designs. *Biometrical Journal.* **46(1)**, 114-124.

Chaudhuri, Arijit, Adhikary, Arun Kumar and Dihidar, Shankar (2000). Mean square error estimation in multi-stage sampling. *Metrika.* **52**, 115-131.

Chaudhuri, Arijit and Dihidar, Kajal (2009). Estimating means of stigmatizing qualitative and quantitative variables from discretionary responses randomized or direct. *Sankhya, Series B.* **71(1)**, 123-136.

Chaudhuri, Arijit and Mukerjee, Rahul (1985). Optionally randomized response techniques. *Calcutta Statistical Association Bulletin.* **34**, 225-229.

Chaudhuri, Arijit and Mukerjee, Rahul (1988). Randomized response: Theory and techniques. Marcel Dekker. New York.

Chaudhuri, Arijit and Saha, Amitava (2005). Optional versus compulsory randomized response techniques in complex surveys. *Journal of Statistical Planning and Inference.* **135**, 516-527.

Eichorn, B.H. and Hayre, L.S. (1983). Scrabmled RR method for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference.* **7**, 307-316.

Gjestvang, C.R. and Singh, S. (2009). An improved randomized response model: Estimation of mean. *Journal of Applied Statistics.* **36(12)**, 1361-1367.

Gupta, Sat, Gupta, Bhisam and Singh, Sarjinder (2002). Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and Inference.* **100**, 239-247.

Himmelfarb, S. and Edgell, S.E. (1980). Additive constant model: A randomized response technique for eliminating evasiveness to quantitative response questions. *Psychological Bulletin.* **87(3)**, 525-530.

Mangat, N.S. and Singh, Ravindra (1990). An alternative randomized response procedure. *Biometrika.* **77**, 439-442.

Mangat, N.S. and Singh, Sarjinder (1994). Optional randomized response model. *Journal of Indian Statistical Association.* **32(3)**, 71-75.

Pal, Sanghamitra (2008). Unbiasedly estimating the total of a stigmatizing variable from a complex survey on permitting options for direct or randomized responses. *Statistical Papers.* **49**, 157-164.

Raj, Des (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association.* **61**, 391-396.

Rao, J.N.K., Hartley, H.O., and Cochran, W.G. (1962).On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society.* **24**, 482-491.

Singh, Sarjinder and Joarder, A.H. (1997). Optional randomized response technique for sensitive quantitative variable. *Metron.* **55**, 151-157.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association.* **60**, 63-69.