Hacettepe Journal of Mathematics and Statistics
Volume 45 (6) (2016), 1819 – 1830

# A two-step approach to ratio and regression estimation of finite population mean using optional randomized response models

Geeta Kalucha*, Sat Gupta† and Javid Shabbir‡

**Abstract**

We propose a modified two-step approach for estimating the mean of a sensitive variable using an additive optional RRT model which allows respondents the option of answering a quantitative sensitive question directly without using the additive scrambling if they find the question non-sensitive. This situation has been handled before in Gupta et al. (2010) using the split sample approach. In this work we avoid the split sample approach which requires larger total sample size. Instead, we estimate the finite population mean by using an Optional Additive Scrambling RRT Model but the corresponding sensitivity level is estimated from the same sample by using the traditional Binary Unrelated Question RRT Model of Greenberg et al. (1969). The initial mean estimation is further improved by utilizing information from a non-sensitive auxiliary variable by way of ratio and regression estimators. Expressions for the *Bias* and *MSE* of the proposed estimators (correct up to first order approximation) are derived. We compare the results of this new model with those of the split-sample based Optional Additive RRT Model of Kalucha et al. (2015), Gupta et al. (2015) and the simple optional additive RRT Model of Gupta et al. (2010). We see that the regression estimator for the new model has the smallest *MSE* among all of the estimators considered here when they have the same sample size.

**Keywords:** Auxiliary Information, Mean square error, Optional randomized response technique, Ratio estimator, Regression estimator, Unrelated Question RRT Model

*2000 AMS Classification:* 62D05

*Department of Mathematics, University of Delhi, India,
Email: geetakalucha@gmail.com

†Department of Mathematics and Statistics, The University of North Carolina at Greensboro, North Carolina, USA,
Email: sngupta@ungc.edu (Corresponding Author.)

‡Department of Statistics, Quaid-I-Azam University, Islamabad, Pakistan,
Email: javidshabbir@gmail.com

*Received* : 17.12.2014 *Accepted* : 27.05.2015 *Doi* : 10.15672/HJMS.20156811029

## 1. Introduction

The randomized response technique of reducing respondent bias in obtaining answers to sensitive questions developed by Warner (1965) has been extended from the situation where response is categorical to that in which the response is quantitative. Choice of scrambling mechanism plays an important role in quantitative response models. Eichhron and Hayre (1983), Gupta and Shabbir (2004), Gupta et al. (2002, 2010), Wu et al. (2008) and many others have estimated the mean of a sensitive variable when the study variable is sensitive and no auxiliary information is available. While Eichhron and Hayre (1983) have used multiplicative scrambling, Gupta et al. (2010) have used additive scrambling in the context of optional randomized response models where a respondent provides a true response if he/she considers the question non-sensitive, and provides a scrambled response if the question is deemed sensitive. The researcher will not know which type of response has been provided. Sousa et al. (2010) and Gupta et al. (2012) suggested mean estimators based on full additive RRT models using an auxiliary variable. Kalucha et al. (2015) and Gupta et al. (2015) improved the mean estimators further by using optional additive RRT models which apart from estimating $\mu_Y$ (the mean of sensitive variable $Y$) also estimated $W$ (the sensitivity level of the research question) using a split-sample approach. Recently Singh and Tarray (2014) have studied optional randomized response model in the stratified sampling setting.

The main motivation for the proposed model is to avoid the split sample approach which requires unnecessarily larger total sample sizes. We estimate the mean of the sensitive characteristic by using an Additive Optional RRT model but the corresponding sensitivity level is estimated from the same sample by using the Greenberg et al. (1969) model. This eliminates the need for split-sample approach that requires a larger total sample size.

Let $\mu_Y$ and $\sigma_Y^2$ be the unknown mean and variance of the sensitive variable $Y$, $\mu_X$ and $\sigma_X^2$ be the known mean and variance of the auxiliary variable $X$. Let $W$ be the unknown sensitivity level of the survey question in the population.

## 2. The Split-Sample Model – Gupta et al. (2010)

Here the sample of size $n$ is split into two sub-samples of sizes $n_1$ and $n_2$ ($n_1+n_2=n$). Let $S_1$, $S_2$ be scrambling variables used in the two sub-samples. Let the mean and variance respectively of $S_i$ $(i=1,2)$ be $\theta_i$ and $\sigma_{S_i}^2$. We assume that $Y$, $X$ and $S_i$ $(i=1,2)$ are mutually independent. For the $i^{\text{th}}$ population unit $(i=1,2,\ldots,N)$, let $y_i$ and $x_i$ respectively be the values of the study variable $Y$ and the auxiliary variable $X$. Moreover let $\bar{y}=\frac{\sum_1^n y_i}{n}$, $\bar{x}=\frac{\sum_1^n x_i}{n}$, $\bar{z}=\frac{\sum_1^n z_i}{n}$ be the sample means, and $\mu_Y=E(Y)$, $\mu_X=E(X)$ and $\mu_Z=E(Z)$ be the corresponding population means for $Y$, $X$ and the scrambled response $Z$ respectively. We assume that $\mu_X$ is known. In each sub sample, we will observe $X$ directly but will only have an additively scrambled version of $Y$. According to this model, the reported response $Z_i$ in the $i^{\text{th}}$ sub-sample is given by

$$Z_i=\left\{\begin{array}{ll} Y & \text{with probability } (1-W) \\ (Y+S_i) & \text{with probability } W \end{array}\right\} \quad i=1,2$$

The mean and variance respectively for $Z_i$ $(i=1,2)$ are given by

$$E(Z_i)=\mu_Y+\theta_i W \quad \text{where } E(S_i)=\theta_i \ (i=1,2), \tag{2.1}$$

and

$$\sigma_{Z_i}^2 = \sigma_Y^2 + \sigma_{S_i}^2 W + \theta_i^2 W(1-W) \tag{2.2}$$

It follows easily from (2.1) that for $\theta_1 \neq \theta_2$,

$$\mu_Y = \frac{\theta_2 E(Z_1) - \theta_1 E(Z_2)}{\theta_2 - \theta_1} \quad \text{and} \quad W = \frac{E(Z_2) - E(Z_1)}{(\theta_2 - \theta_1)}\,. \tag{2.3}$$

Hence if information on $X$ is ignored, expressions in (2.3) lead to the following unbiased estimators of $\mu_Y$ and $W$:

$$\hat{\mu}_Y = \frac{\theta_2 \bar{z}_1 - \theta_1 \bar{z}_2}{\theta_2 - \theta_1},\ \theta_1 \neq \theta_2 \quad \text{and} \quad \hat{W} = \frac{\bar{z}_2 - \bar{z}_1}{(\theta_2 - \theta_1)},\ \theta_1 \neq \theta_2, \tag{2.4}$$

where $\bar{z}_1,\ \bar{z}_2$ respectively are the sample mean of reported responses in the two subsamples.

It can be verified that $\hat{\mu}_Y$ and $\hat{W}$ are unbiased estimators of the population mean $\mu_Y$ and the sensitivity level $W$. Variances of these estimators are given by

$$\mathrm{Var}(\hat{\mu}_Y) = \frac{1}{(\theta_2 - \theta_1)^2}\left[\theta_2^2\left(\frac{1-f_1}{n_1}\right)\sigma_{Z_1}^2 + \theta_1^2\left(\frac{1-f_2}{n_2}\right)\sigma_{Z_2}^2\right] \tag{2.5}$$

and

$$\mathrm{Var}(\hat{W}) = \frac{1}{(\theta_2 - \theta_1)^2}\left[\left(\frac{1-f_1}{n_1}\right)\sigma_{Z_1}^2 + \left(\frac{1-f_2}{n_2}\right)\sigma_{Z_2}^2\right],$$

where $\theta_1 \neq \theta_2$, $f_1 = \frac{n_1}{N}$, $f_2 = \frac{n_2}{N}$, $f = \frac{n}{N} = f_1 + f_2$,

$$\sigma_{Z_1}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Z_{1_i} - \mu_Z)^2 \quad \text{and} \quad \sigma_{Z_2}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Z_{2_i} - \mu_Z)^2\,.$$

## 3. The Proposed Model

In the proposed model, the underlying sensitivity level $W$ and its variance are estimated by using the Greenberg et al. (1969) model. Here the sensitive question is "Whether or not you consider the underlying main research question sensitive for a face-to-face survey". Let $\pi_b$ be the known probability of the binary innocuous unrelated question and $p_b$ be the known probability of the respondent selecting the sensitivity question. We consider a finite population $U = \{1, 2, \ldots, N\}$ of size $N$ and a random sample of size $n$ be drawn without replacement. When estimating the mean, let $S$ be the scrambling variable used to additively scramble the responses in the sample with mean $E(S) = \theta$. We assume that $Y$, $X$ and $S$ are mutually independent.

**3.1. Estimation of Sensitivity Level** $(W)$**.** The probability of "yes response" to the sensitivity question is given by

$$P_y = p_b W + (1 - p_b)\pi_b \tag{3.1}$$

Solving for $W$, we have

$$W = \frac{P_y - (1 - p_b)\pi_b}{p_b} \tag{3.2}$$

Thus the estimate of $W$, as per the Greenberg et al. (1969) model, is given by

$$\hat{W} = \frac{\hat{P}_y - (1 - p_b)\pi_b}{p_b}, \tag{3.3}$$

where $\hat{P}_y$ is the proportion of yes response in the sample.

We know that $\hat{W}$ is an unbiased estimator and its variance is given by

$$\text{Var}(\hat{W}) = \left(\frac{1-f}{n}\right)\frac{P_y(1-P_y)}{p_b^2} \tag{3.4}$$

An unbiased estimator of this variance is given by

$$\hat{\text{Var}}(\hat{W}) = \left(\frac{1-f}{n-1}\right)\frac{\hat{P}_y(1-\hat{P}_y)}{p_b^2} \tag{3.5}$$

**3.2. Estimation of Mean.** The reported quantitative response $Z$ to the main research question according to optional additive RRT model can be expressed as

$$Z = \left\{ \begin{array}{ll} Y+S & \text{with probability } W \\ Y & \text{with probability } 1-W \end{array} \right\}$$

The mean and variance respectively of $Z$ are given by

$$\begin{aligned} E(Z) &= WE(Y+S) + (1-W)E(Y) \\ &= E(Y) + WE(S) \\ &= \mu_Y + W\theta, \end{aligned} \tag{3.6}$$

and

$$\begin{aligned} \text{Var}(Z) &= WE(Y+S)^2 + (1-W)E(Y^2) - \mu_Z^2 \\ &= \sigma_Y^2 + W\sigma_S^2 + \theta^2 W(1-W) \end{aligned} \tag{3.7}$$

From equation (3.6) we have

$$\mu_Y = \mu_Z - W\theta$$

This leads to an estimator for $\mu_Y$ given by

$$\hat{\mu}_{YW^*} = \hat{\mu}_Z - \hat{W}\theta, \tag{3.8}$$

where $\hat{\mu}_Z = \bar{z}$ is the sample mean of reported responses and $\hat{W}$ is given by equation (3.3).

We note that $\hat{\mu}_{YW^*}$ is an unbiased estimator of $\mu_Y$ and its variance is given by

$$\begin{aligned} \text{Var}(\hat{\mu}_{YW^*}) &= \text{Var}(\bar{z} - \hat{W}\theta) \\ &= \text{Var}(\bar{z}) + \theta^2 \,\text{Var}(\hat{W}) \\ &= \left(\frac{1-f}{n}\right)(\sigma_Z^2) + \theta^2\left(\frac{1-f}{n}\right)\frac{P_y(1-P_y)}{p_b^2} \end{aligned} \tag{3.9}$$

The variance of the estimator in (3.9) can be conveniently estimated by

$$\hat{\text{Var}}(\hat{\mu}_{YW^*}) = \left(\frac{1-f}{n}\right)(s_z^2) + \theta^2\hat{\text{Var}}(\hat{W}) \tag{3.10}$$

where $s_z^2$ is the sample variance of reported responses given by
$s_z^2 = (n-1)^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$ and $\hat{\text{Var}}(\hat{W})$ is as given in (3.5) above.

We further modify the proposed mean estimator $\hat{\mu}_{YW^*}$ in the presence of an auxiliary variable by proposing ratio ($\hat{\mu}_{RW^*}$) and regression ($\hat{\mu}_{RegW^*}$) estimators and compare it with the estimators proposed in Kalucha et al. (2015) and Gupta et al. (2015), both based on split-sample approach.

## 4. Ratio Estimator

**4.1. Kalucha et al. (2015) – Split-Sample Based Ratio Estimator.** Kalucha et al. (2015) proposed the following additive ratio estimator for the mean of $Y$:

$$\hat{\mu}_{AR} = \left(\frac{\theta_2\bar{z}_1 - \theta_1\bar{z}_2}{\theta_2 - \theta_1}\right)\left(\frac{\mu_X}{\bar{x}_1} + \frac{\mu_X}{\bar{x}_2}\right)\left(\frac{1}{2}\right), \quad \theta_1 \neq \theta_2. \tag{4.1}$$

where $\left(\frac{\theta_2\bar{z}_1 - \theta_1\bar{z}_2}{\theta_2 - \theta_1}\right)$ is the unbiased estimator of $\mu_Y$ given by Gupta et al. (2010), and $\bar{x}_1$ and $\bar{x}_2$ are the respective sub-sample means for $X$. It was shown that this estimator performs better than the ratio estimator proposed by Sousa et al. (2010) utilizing a non-optional additive RRT model.

*Bias* and *MSE* of $\hat{\mu}_{AR}$, correct up to first order of approximation, are given by

$$\begin{aligned} Bias(\hat{\mu}_{AR}) &= \left(\frac{1-f_1}{n_1}\right)\left[\frac{\mu_Y}{2}C_x^2 - \left(\frac{\theta_2}{\theta_2-\theta_1}\right)\frac{\rho_{yx}\sigma_Y C_x}{2}\right] \\ &\quad + \left(\frac{1-f_2}{n_2}\right)\left[\frac{\mu_Y}{2}C_x^2 + (\frac{\theta_1}{\theta_2-\theta_1})\frac{\rho_{yx}\sigma_Y C_x}{2}\right] \\ &= C_x^2\mu_Y\left[\alpha - \rho_{yx}\frac{\beta}{2}\right] \end{aligned} \tag{4.2}$$

and

$$\begin{aligned} MSE(\hat{\mu}_{AR}) &= \frac{1}{(\theta_2-\theta_1)^2}\left[\theta_2^2\left(\frac{1-f_1}{n_1}\right)\sigma_{Z_1}^2 + \theta_1^2\left(\frac{1-f_2}{n_2}\right)\sigma_{Z_2}^2\right] \\ &\quad + \frac{\mu_Y^2 C_x^2}{4}\alpha - \mu_Y\rho_{yx}\sigma_Y C_x\beta \end{aligned} \tag{4.3}$$

where $\alpha = \left(\frac{1-f_1}{n_1}\right) + \left(\frac{1-f_2}{n_2}\right)$, $\beta = \left(\frac{1-f_1}{n_1}\right)\left(\frac{\theta_2}{\theta_2-\theta_1}\right) - \left(\frac{1-f_2}{n_2}\right)\left(\frac{\theta_1}{\theta_2-\theta_1}\right)$, and $C_x$ is the coefficient of variation for $X$.

**4.2. Proposed Ratio Estimator-New Approach.** In this section we propose a ratio estimator where the RRT estimator of the mean of $Y$ given by (3.8) above is further improved by using information on an auxiliary variable $X$. We define $\delta_z = (\bar{z} - \mu_Z)/\mu_Z$, $\delta_x = (\bar{x} - \mu_X)/\mu_X$. Note that $E(\delta_i) = 0$ for $i = z, x$.

The proposed estimator is given by

$$\hat{\mu}_{RW^*} = (\bar{z} - \hat{W}\theta)\left(\frac{\mu_X}{\bar{x}}\right) = (\mu_Z(1+\delta_z) - \hat{W}\theta)(1+\delta x)^{-1} \tag{4.4}$$

Using Taylor's approximation and retaining terms of order up to 2, (4.4) can be rewritten as

$$\hat{\mu}_{RW^*} - \mu_Z \cong \mu_Z(\delta_z - \delta_x - \delta_z\delta_x + \delta_x^2) - \hat{W}\theta(1 - \delta_x + \delta_x^2) \tag{4.5}$$

Substituting the value of $\mu_Z$ from (3.6) in (4.5), we have

$$\hat{\mu}_{RW^*} - \mu_Y \cong \mu_Y(\delta_z - \delta_x - \delta_z\delta_x + \delta_x^2) + (W - \hat{W})\theta(1 - \delta_x + \delta_x^2) + W\theta(\delta_z - \delta_z\delta_x) \tag{4.6}$$

Under the assumption of bivariate normality (see Sukhatme and Sukhatme, 1970), we have

$$E(\delta_z^2) = \frac{1-f}{n}C_z^2,\ E(\delta_x^2) = \frac{1-f}{n}C_x^2,\ E(\delta_z\delta_x) = \frac{1-f}{n}C_{zx}$$

where $C_{zx} = \rho_{zx}C_zC_x$, $C_z$ and $C_x$ are the coefficients of variation of $Z$ and $X$, respectively.

Also, we have:

$$C_z^2 = \frac{\sigma_y^2 + W\sigma_S^2 + \theta^2 W(1-W)}{(\bar{Z})^2} \quad \text{and} \quad \rho_{zx} = \frac{\rho_{yx}}{\sqrt{1 + W\frac{\sigma_S^2}{\sigma_y^2} + \frac{\theta^2 W(1-W)}{\sigma_y^2}}} \tag{4.7}$$

From equation (4.6), we can get expression for the *Bias* of $\hat{\mu}_{RW^*}$, correct up to first order of approximation, as given by

$$Bias(\hat{\mu}_{RW^*}) \cong \mu_Y \left(\frac{1-f}{n}\right)(C_x^2 - \rho_{zx}C_zC_x) - W\theta\left(\frac{1-f}{n}\right)\rho_{zx}C_zC_x \tag{4.8}$$

Similarly from (4.6), *MSE* of $\hat{\mu}_{RW^*}$, correct to first order of approximation, is given by

$$\begin{aligned} MSE(\hat{\mu}_{RW^*}) &= E(\hat{\mu}_{RW^*} - \mu_Y)^2 \\ &\cong \mu_Y^2 E(\delta_z^2 + \delta_x^2 - 2\delta_z\delta_x) + \theta^2 E(W - \hat{W})^2 E(1 - 2\delta_x + 3\delta_x^2) \\ &\quad + W^2\theta^2 E(\delta_z^2) + 2\mu_Y W\theta E(\delta_z^2 - \delta_z\delta_x) \end{aligned}$$

or

$$\begin{aligned} MSE(\hat{\mu}_{RW^*}) \cong{}& \left(\frac{1-f}{n}\right)\mu_Y^2(C_z^2 + C_x^2 - 2\rho_{zx}C_zC_x) \\ &+ \theta^2 \operatorname{Var}(\hat{W})\left(1 + 3\left(\frac{1-f}{n}\right)C_x^2\right) + W^2\theta^2\left(\frac{1-f}{n}\right)C_z^2 \\ &+ 2\mu_Y W\theta\left(\frac{1-f}{n}\right)(C_z^2 - \rho_{zx}C_zC_x) \end{aligned} \tag{4.9}$$

where $\operatorname{Var}(\hat{W})$ is given by (3.4) above.

**4.3. Mean and Variance of the Proposed Ratio Estimator.** The proposed ratio estimator can be rewritten as

$$\hat{\mu}_{RW^*} = \left(\frac{\bar{y}}{\bar{x}}\right)\mu_X, \quad \text{where } \bar{y} = \bar{z} - \hat{W}\theta \tag{4.10}$$

Hence

$$E(\hat{\mu}_{RW^*}) = \mu_X E\left\{\frac{\bar{y}}{\bar{x}}\right\} \tag{4.11}$$

Using a Taylor series expansion of $\frac{\bar{y}}{\bar{x}}$ around $(\mu_Y,\mu_X)$:

$$\begin{aligned} \frac{\bar{y}}{\bar{x}} \cong{}& \frac{\vec{y}}{\bar{x}}\bigg|_{(\mu_Y,\mu_X)} + (\bar{y} - \mu_Y)\frac{\partial}{\partial\bar{y}}\left(\frac{\bar{y}}{\bar{x}}\right)\bigg|_{(\mu_Y,\mu_X)} + (\bar{x} - \mu_X)\frac{\partial}{\partial\bar{x}}\left(\frac{\bar{y}}{\bar{x}}\right)\bigg|_{(\mu_Y,\mu_X)} \\ &+ \frac{1}{2}(\bar{y} - \mu_Y)^2\frac{\partial^2}{\partial\bar{y}^2}\left(\frac{\bar{y}}{\bar{x}}\right)\bigg|_{(\mu_Y,\mu_X)} + \frac{1}{2}(\bar{x} - \mu_X)^2\frac{\partial^2}{\partial\bar{x}^2}\left(\frac{\bar{y}}{\bar{x}}\right)\bigg|_{(\mu_Y,\mu_X)} \\ &+ (\bar{y} - \mu_Y)(\bar{x} - \mu_X)\frac{\partial^2}{\partial\bar{y}\partial\bar{x}}\left(\frac{\bar{y}}{\bar{x}}\right)\bigg|_{(\mu_Y,\mu_X)} \\ &+ O\left(\left((\bar{y} - \mu_Y)\frac{\partial}{\partial\bar{y}} + (\bar{x} - \mu_X)\frac{\partial}{\partial\bar{x}}\right)^3\left(\frac{\bar{y}}{\bar{x}}\right)\right) \end{aligned}$$

The mean of $\frac{\bar{y}}{\bar{x}}$ can now be found by taking expected value, ignoring all terms higher than 2.

$$\begin{aligned} E\left\{\frac{\bar{y}}{\bar{x}}\right\} &\cong \frac{\mu_Y}{\mu_X} + \operatorname{Var}(\bar{x})\frac{\mu_Y}{\mu_X^3} - \frac{\operatorname{Cov}(\bar{y},\bar{x})}{\mu_X^2} \\ &\cong \frac{\mu_Y}{\mu_X} + \frac{(1-f)}{n}\left(\operatorname{Var}(x)\frac{\mu_Y}{\mu_X^3} - \frac{\operatorname{Cov}(y,x)}{\mu_X^2}\right) \end{aligned} \tag{4.12}$$

Substituting (4.12) in (4.11), we get

$$E(\hat{\mu}_{RW^*}) \cong \mu_Y + \frac{(1-f)}{n}\left(\mathrm{Var}(x)\frac{\mu_Y}{\mu_X^2} - \frac{\mathrm{Cov}(y,x)}{\mu_X}\right) \tag{4.13}$$

It is clear from the above expression that $\hat{\mu}_{RW^*}$ is asymptotically unbiased. Now

$$\mathrm{Var}(\hat{\mu}_{RW^*}) = \mu_X^2\,\mathrm{Var}\left(\frac{\bar{y}}{\bar{x}}\right) \tag{4.14}$$

An approximation of the variance of $\frac{\bar{y}}{\bar{x}}$ is obtained by using the first order terms of Taylor series expansion:

$$\begin{aligned}
\mathrm{Var}\left(\frac{\bar{y}}{\bar{x}}\right) &= E\left\{\left(\frac{\bar{y}}{\bar{x}} - E\left\{\frac{\bar{y}}{\bar{x}}\right\}\right)^2\right\} \\
&\cong E\left\{\left(\frac{\bar{y}}{\bar{x}} - \frac{\mu_Y}{\mu_X}\right)^2\right\} \\
&\cong \frac{\mathrm{Var}(\bar{y})}{\mu_X^2} + \frac{\mu_Y^2\,\mathrm{Var}(\bar{x})}{\mu_X^4} - \frac{2\mu_Y\,\mathrm{Cov}(\bar{y},\bar{x})}{\mu_X^3} \\
&\cong \frac{(1-f)}{n}\left(\frac{\mathrm{Var}(y)}{\mu_X^2} + \frac{\mu_Y^2\,\mathrm{Var}(x)}{\mu_X^4} - \frac{2\mu_Y\,\mathrm{Cov}(y,x)}{\mu_X^3}\right) \qquad (4.15)
\end{aligned}$$

Substituting (4.15) in (4.14), we have

$$\mathrm{Var}(\hat{\mu}_{RW^*}) \cong \frac{(1-f)}{n}\left(\mathrm{Var}(y) + \frac{\mu_Y^2\,\mathrm{Var}(x)}{\mu_X^2} - \frac{2\mu_Y\,\mathrm{Cov}(y,x)}{\mu_X}\right) \tag{4.16}$$

Substituting for $\mathrm{Var}(y)$ and using the fact that $\mathrm{Cov}(y,x) = \mathrm{Cov}(z,x)$ in (4.16), we get

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{RW^*}) \cong \frac{(1-f)}{n}\Bigg(&\mathrm{Var}(z) - W\,\mathrm{Var}(S) - \theta^2 W(1-W) \\
&+ \frac{\mu_Y^2\,\mathrm{Var}(x)}{\mu_X^2} - \frac{2\mu_Y\,\mathrm{Cov}(z,x)}{\mu_X}\Bigg)
\end{aligned} \tag{4.17}$$

The above variance can be estimated by using:

$$\hat{\mathrm{Var}}(z) = s_z^2, \quad \hat{W} = \frac{\hat{P}_y - (1-p_b)\pi_b}{p_b}, \text{ and } \hat{\mathrm{Cov}}(z,x) = s_{zx},$$

where sample covariance $s_{zx} = (n-1)^{-1}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})$.

## 5. Regression Estimator

**5.1. Gupta et al. (2015) – Split-Sample Based Regression Estimator.** Gupta et al. (2015) suggested a regression estimator of the mean using split-sample approach, as given by:

$$\hat{\mu}_{Areg} = \left(\frac{\theta_2\bar{z}_1 - \theta_1\bar{z}_2}{\theta_2 - \theta_1}\right) + \left\{\hat{\beta}_{Z_1X_1}(\mu_X - \bar{x}_1) + \hat{\beta}_{Z_2X_2}(\mu_X - \bar{x}_2)\right\}\left(\frac{1}{2}\right), \tag{5.1}$$

where $\hat{\beta}_{Z_iX_i}(i = 1,2)$ are the sample regression coefficients between $Z_i$ and $X_i$ respectively, and $\bar{z}_i$, $\bar{x}_i$ $(i = 1,2)$ are the two sub-sample means. It was shown that this estimator performs better than the regression estimator proposed by Gupta et al. (2012) utilizing a non-optional additive RRT model. *Bias* and *MSE* of $\hat{\mu}_{Areg}$, correct up to first order of approximation, are given by

$$Bias(\hat{\mu}_{Areg}) \cong \left[-\frac{1}{2}\beta_{Z_1X}\left(\frac{1-f_1}{n_1}\right) - \frac{1}{2}\beta_{Z_2X}\left(\frac{1-f_2}{n_2}\right)\right]\left\{\frac{\mu_{12}}{\mu_{11}} - \frac{\mu_{03}}{\mu_{02}}\right\} \tag{5.2}$$

and

$$MSE^{(1)}(\hat{\mu}_{A\,reg}) = \frac{1}{(\theta_2-\theta_1)^2}\left[\theta_2^2\left(\frac{1-f_1}{n_1}\right)\sigma_{Z_1}^2 + \theta_1^2\left(\frac{1-f_2}{n_2}\right)\sigma_{Z_2}^2\right] + \frac{\rho_{yx}^2\sigma_Y^2}{4}\alpha - \rho_{yx}^2\sigma_Y^2\beta, \tag{5.3}$$

where $\theta_2 \neq \theta_1$; $\alpha$ and $\beta$ are defined earlier and $\mu_{rs} = \frac{1}{N-1}\sum_{i=1}^{N}(z_i-\bar{Z})^r(x_i-\bar{X})^s$.

**5.2. Proposed Regression Estimator-New Approach.** We modify the mean estimator in (3.8) above by using the regression estimation approach and propose the following estimator for the population mean of $Y$:

$$\hat{\mu}_{Reg\,W^*} = (\bar{z} - \hat{W}\theta) + \hat{\beta}_{zx}(\mu_X - \bar{x}) \tag{5.4}$$

We obtain the expressions for the bias and the mean square error for the proposed regression estimator $\hat{\mu}_{Reg\,W^*}$. If $e_0 = (\bar{z}-\mu_Z)/\mu_Z$, $e_1 = (\bar{x}-\mu_X)/\mu_X$, $e_2 = (\sigma_x^2-\sigma_X^2)/\sigma_X^2$ and $e_3 = (\sigma_{zx}-\sigma_{ZX})/\sigma_{ZX}$, then we have $E(e_i) = 0$, $i = 0, 1, 2, 3$.

Using Taylor's approximation and retaining terms of order up to 2, (5.4) can be rewritten as

$$\hat{\mu}_{Reg\,W^*} - \mu_Z \cong \mu_Z e_0 - \hat{W}\theta - \beta_{zx}\mu_X[e_1 + e_1e_3 - e_1e_2] \tag{5.5}$$

Substituting for $\mu_Z$, (5.5) can be written as

$$\hat{\mu}_{Reg\,W} - \mu_Y \cong \mu_Z e_0 - \beta_{zx}\mu_X[e_1 + e_1e_3 - e_1e_2] + (W - \hat{W})\theta \tag{5.6}$$

From Mukhopadhyay (1998, p. 123), we have $E(e_1^2) = \frac{1-f}{n}C_x^2$, $E(e_0^2) = \frac{1-f}{n}C_z^2$, $E(e_1e_2) = \frac{1-f}{n}\frac{1}{\bar{X}}\frac{\mu_{03}}{\mu_{02}}$, $E(e_1e_3) = \frac{1-f}{n}\frac{1}{\bar{X}}\frac{\mu_{12}}{\mu_{11}}$, where $\mu_{rs} = \frac{1}{N-1}\sum_{i=1}^{N}(z_i-\bar{Z})^r(x_i-\bar{X})^s$ and $C_x$, $C_z$ are the coefficients of variation of $x$ and $z$, respectively. Also, we have:

$$\beta_{zx} = \frac{\sigma_{zx}}{\sigma_x^2} = \frac{\sigma_{yx}}{\sigma_x^2} = \rho_{yx}\frac{\sigma_y}{\sigma_x} = \beta_{yx} \tag{5.7}$$

where $\rho_{yx}$ and $\rho_{zx}$ are the coefficients of correlation between $y$ and $x$, and between $z$ and $x$, respectively.

Using this in (5.6), the *Bias* of $\hat{\mu}_{Reg\,W}$, to first order of approximation, is given by

$$Bias(\hat{\mu}_{Reg\,W^*}) \cong -\beta_{zx}\left(\frac{1-f}{n}\right)\left\{\frac{\mu_{12}}{\mu_{11}} - \frac{\mu_{03}}{\mu_{02}}\right\} \tag{5.8}$$

The expression for *MSE* of $\hat{\mu}_{Reg\,W^*}$ to first order of approximation, is given by

$$\begin{aligned} MSE(\hat{\mu}_{Reg\,W^*}) &\cong \left(\frac{1-f}{n}\right)\left[\sigma_z^2 - \frac{\sigma_{yx}^2}{\sigma_x^2}\right] + \theta^2\,\mathrm{Var}(\hat{W}) \\ &= \left(\frac{1-f}{n}\right)\sigma_y^2\left\{\left(1 + \frac{W\sigma_S^2 + \theta^2 W(1-W)}{\sigma_y^2}\right) - \rho_{yx}^2\right\} + \theta^2\,\mathrm{Var}(\hat{W}) \end{aligned} \tag{5.9}$$

where $\mathrm{Var}(\hat{W})$ is given by (3.4) above.

We note that $\hat{\mu}_{Reg\,W^*}$ is an unbiased estimator and hence

$$\begin{aligned} \mathrm{Var}(\hat{\mu}_{Reg\,W^*}) &= MSE(\hat{\mu}_{Reg\,W^*}) \\ &\cong \left(\frac{1-f}{n}\right)\left[\sigma_z^2 - \frac{\sigma_{yx}^2}{\sigma_x^2}\right] + \theta^2\,\mathrm{Var}(\hat{W}) \end{aligned} \tag{5.10}$$

The above variance can be estimated by using:

$$\hat{\sigma}_z^2 = s_z^2,\ \hat{\sigma}_{yx}^2 = \hat{\sigma}_{zx}^2 = s_{zx}^2 \quad \text{and} \quad \hat{\mathrm{Var}}(\hat{W}) = \frac{(1-f)}{(n-1)}\frac{\hat{P}_y(1-\hat{P}_y)}{p_b^2}.$$

## 6. Efficiency comparisons

**6.1. Efficiency Comparison of $\hat{\mu}_{RW^*}$ and $\hat{\mu}_{YW^*}$.** We have from equations (3.9) and (4.9), $MSE(\hat{\mu}_{RW^*}) < MSE(\hat{\mu}_{YW^*})$ if

$$1 + \frac{3\theta^2 \operatorname{Var}(\hat{W})}{\mu_Y^2} < 2\rho_{yx}\frac{C_y}{C_x} \tag{6.1}$$

Since $\frac{3\theta^2 \operatorname{Var}(\hat{W})}{\mu_Y^2}$ approaches 0 because $\operatorname{Var}(\hat{W})$ approaches 0 as the sample becomes larger, (6.1) will generally hold if

$$1 < 2\rho_{yx}\frac{C_y}{C_x} \text{ or } \rho_{yx} > \frac{1}{2}\frac{C_x}{C_y} \tag{6.2}$$

If we assume ($C_x \approx C_y$), we can conclude from (6.2) that

$$MSE(\hat{\mu}_{RW^*}) < MSE(\hat{\mu}_{YW^*}) \text{ if } \rho_{yx} > \frac{1}{2}. \tag{6.3}$$

Hence the proposed ratio estimator ($\hat{\mu}_{RW^*}$) is more efficient than the proposed ordinary mean estimator ($\hat{\mu}_{YW^*}$) when the correlation between the study variable and the auxiliary variable is high $\left(\rho_{yx} > \frac{1}{2}\right)$.

**6.2. Efficiency Comparison of $\hat{\mu}_{RegW^*}$ with $\hat{\mu}_{RW^*}$ and $\hat{\mu}_{YW^*}$.**

(i) It can be verified from (3.9) and (5.9) that according to first order approximation $MSE(\hat{\mu}_{RegW^*}) < MSE(\hat{\mu}_{YW^*})$ if

$$\left(\frac{1-f}{n}\right)\frac{\sigma_{yx}^2}{\sigma_x^2} > 0 \tag{6.4}$$

(ii) It can be verified from (4.9) and (5.9) that up to first order approximation $MSE(\hat{\mu}_{RegW^*}) < MSE(\hat{\mu}_{RW^*})$ if

$$1 - 2\rho_{yx}\frac{C_y}{C_x} + \rho_{yx}^2\frac{C_y^2}{C_x^2} + \frac{3\theta^2 \operatorname{Var}(\hat{W})}{\mu_Y^2} > 0 \tag{6.5}$$

With ($C_x \cong C_y$), (6.5) can be rewritten as

$$(1-\rho_{yx})^2 + \frac{3\theta^2 \operatorname{Var}(\hat{W})}{\mu_Y^2} > 0 \tag{6.6}$$

Since the conditions (6.4) and (6.6) will always hold true, up to first order of approximation, the regression estimator $\hat{\mu}_{RegW^*}$ performs better than the ordinary mean estimator $\hat{\mu}_{YW^*}$ and the ratio estimator $\hat{\mu}_{RW^*}$.

## 7. Simulation Study

**7.1. Comparison of the Proposed Model with the Split-Sample Model in the Presence of Auxiliary Information.** The tables below provide a comparison between the proposed model and the split-sample additive scrambling models of Kalucha et al. (2015) and Gupta et al. (2015) in the presence of non-sensitive auxiliary information. We choose the parameters as per the observation A1 (given below) that was obtained in Gupta et al. (2015) under which the regression estimator $\hat{\mu}_{Areg}$ is more efficient than both additive ratio estimator $\hat{\mu}_{AR}$ and the ordinary mean estimator $\hat{\mu}_Y$ under the split sample approach:

**A1**. We choose our scrambling variables $S_1$ and $S_1$ in such a way that their means $\theta_1$ and $\theta_2$ are opposite in signs and associate the one with the smaller magnitude to the larger sub-sample and vice-versa. Also if one of the chosen means is zero then we associate it to the larger split sample.

In the simulation study, we consider a finite population of size $N = 5000$ generated from a bivariate normal distribution. The simulated bivariate normal population has theoretical mean of $[Y, X]$ as $\mu = [6, 4]$. The covariance matrix $(\sum)$ is as given below:

$$\sum = \begin{bmatrix} 9 & 4.8 \\ 4.8 & 4 \end{bmatrix}, \quad \rho_{YX} = 0.7996$$

We estimate the empirical *MSE* using 5000 samples of various sizes selected from this population. The scrambling variables $S_1$ and $S_2$ are taken to be normal variates with $\sigma^2_{S_1} = 2$ and $\sigma^2_{S_2} = 1$. The scrambling variable means are chosen as per **A1 (given above).** The selected means are $\theta_1 = 5$, $\theta_2 = -0.5$ and $n_2 > n_1$. For the population we consider two sample sizes: $n = 500, 1000$ for different values of the sensitivity level $W = 0.3, 0.7, 0.9$.

For the proposed model we choose $\theta = \theta_2 = -0.5$ with $\pi_b = 0.25$ and $p_b = 0.7$.

**Table 1.** Theoretical **(bold)** and empirical *MSE* comparisons of the mean estimator $(\hat{\mu}_{YW^*})$, the ratio estimator $(\hat{\mu}_{RW^*})$ and the regression estimator $(\hat{\mu}_{Reg\,W^*})$ of the proposed model with the mean estimator $(\hat{\mu}_Y)$, the additive ratio estimator $(\hat{\mu}_{AR})$ and the regression estimator $(\hat{\mu}_{Areg})$ of the split-sample model with $\rho_{YX} = 0.7996$.

| $n$ | $W$ | *MSE* Estimation | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Proposed Model | | | | Split-Sample Model | | | | | |
| | | $\mathrm{Var}(\hat{W})$ | $MSE(\hat{\mu}_{YW^*})$ | $MSE(\hat{\mu}_{RW^*})$ | $MSE(\hat{\mu}_{Reg\,W^*})$ | $n_1$ | $n_2$ | $\mathrm{Var}(\hat{W})$ | $MSE(\hat{\mu}_Y)$ | $MSE(\hat{\mu}_{AR})$ | $MSE(\hat{\mu}_{Areg})$ |
| 500 | 0.3 | **0.000749** | **0.017141** | **0.007283** | **0.006706** | 200 | 300 | **0.003511** | **0.024982** | **0.019001** | **0.017437** |
| | | 0.000821 | 0.016916 | 0.007221 | 0.006638 | | | 0.004487 | 0.023217 | 0.018106 | 0.01665 |
| | 0.7 | **0.000903** | **0.0179** | **0.008041** | **0.007465** | 200 | 300 | **0.003688** | **0.02605** | **0.020069** | **0.018505** |
| | | 0.000999 | 0.017614 | 0.008264 | 0.007608 | | | 0.004821 | 0.025906 | 0.020948 | 0.019584 |
| | 0.9 | **0.000764** | **0.018171** | **0.008313** | **0.007736** | 200 | 300 | **0.003277** | **0.026387** | **0.020406** | **0.018842** |
| | | 0.000853 | 0.018221 | 0.008534 | 0.008002 | | | 0.002443 | 0.029625 | 0.023441 | 0.022628 |
| 1000 | 0.3 | **0.000333** | **0.007618** | **0.003237** | **0.002981** | 450 | 550 | **0.001665** | **0.012846** | **0.009044** | **0.008528** |
| | | 0.000416 | 0.00738 | 0.003224 | 0.002915 | | | 0.003114 | 0.011986 | 0.009241 | 0.008602 |
| | 0.7 | **0.000401** | **0.007956** | **0.003574** | **0.003318** | 450 | 550 | **0.001748** | **0.013394** | **0.009593** | **0.009076** |
| | | 0.000497 | 0.007744 | 0.003589 | 0.003319 | | | 0.002965 | 0.012035 | 0.009007 | 0.008506 |
| | 0.9 | **0.000340** | **0.008076** | **0.003694** | **0.003438** | 450 | 550 | **0.001568** | **0.013578** | **0.009777** | **0.009260** |
| | | 0.000423 | 0.008367 | 0.003914 | 0.003693 | | | 0.001270 | 0.012051 | 0.008914 | 0.008395 |

We note from the table that consistently the regression estimator $(\hat{\mu}_{Reg\,W^*})$ is more efficient than the ratio $(\hat{\mu}_{RW^*})$ and the mean estimator $(\hat{\mu}_{YW^*})$ of the proposed model for all values of $W$. Also as the sensitivity $W$ increases, the *MSE*'s increase, highlighting the usefulness of an Optional RRT model since $W$ is highest (equal to 1) for non-optional model. While comparing the proposed model with the split-sample model, we note that *MSE*'s of the proposed model estimators $(\hat{\mu}_{YW^*}, \mu_{RW^*}\mu_{Reg\,W^*})$ are consistently smaller as compared to $(\hat{\mu}_Y, \hat{\mu}_{AR}, \hat{\mu}_{Areg})$ estimators. We observe that for a fixed sample size the *MSE*'s for the proposed model are reduced by more than two and a half times as compared to the split-sample based model.

### 7.2. Comparison of the Point Estimates of Proposed Model with the Split-Sample Model in the Presence of Auxiliary Information.

**Table 2.** Empirical values of the estimators $\hat{W}$, the mean estimator ($\hat{\mu}_{YW^*}$), the ratio estimator ($\hat{\mu}_{RW^*}$) and the regression estimator ($\hat{\mu}_{RegW^*}$) of the proposed model and the corresponding split sample model for $W = 0.3, 0.7, 0.9$ and the population mean $\mu_Y = 6$.

| $n$ | $W$ | Point Estimates | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Proposed Model | | | | Split-sample Model | | | |
| | | $\hat{W}$ | $\hat{\mu}_{YW^*}$ | $\hat{\mu}_{RW^*}$ | $\hat{\mu}_{RegW^*}$ | $\hat{W}$ | $\hat{\mu}_Y$ | $\hat{\mu}_{AR}$ | $\hat{\mu}_{Areg}$ |
| 500 | 0.3 | 0.30049 | 5.91234 | 5.90924 | 5.90958 | 0.34439 | 5.90478 | 5.90812 | 5.90471 |
| | 0.7 | 0.69978 | 5.90947 | 5.91254 | 5.91158 | 0.6523 | 5.86084 | 5.86545 | 5.86143 |
| | 0.9 | 0.89957 | 5.91218 | 5.90169 | 5.91065 | 0.90461 | 5.83561 | 5.83925 | 5.83557 |
| 1000 | 0.3 | 0.30052 | 5.91076 | 5.912 | 5.91161 | 0.34351 | 5.92844 | 5.93066 | 5.92885 |
| | 0.7 | 0.69979 | 5.9116 | 5.91047 | 5.91053 | 0.65809 | 5.89841 | 5.90048 | 5.8986 |
| | 0.9 | 0.89997 | 5.91107 | 5.91144 | 5.91125 | 0.90812 | 5.89409 | 5.89618 | 5.89436 |

We note that both methods produce nearly unbiased estimators of the population mean. However, the proposed model produces better estimates of the sensitivity level.

## Acknowledgements

We would like to thank the two anonymous referees for their careful reading of the paper and for their helpful suggestions which helped improve the presentation.

## References

[1] Eichhron, B.H. and Hayre, L.S. *Scrambled randomized response methods for obtaining sensitive quantitative data*, Journal of Statistical Planning and Inference **7**, 307–316, 1983.

[2] Greenberg, B.G., Abul-Ela, A.L.A., Simmons, W.R. and Horvitz, D.G. *The unrelated question randomized response model-theoretical framework*, Journal of the American Statistical Association **64** (326), 520–539, 1969.

[3] Gupta, S., Kalucha, G. and Shabbir, J. *A regression estimator for finite population mean of a sensitive variable using an optional randomized response model*, Accepted for publication in Communications in Statistics - Simulation and Computation, 2015.

[4] Gupta, S. and Shabbir, J. *Sensitivity estimation for personal interview survey questions*, Statistica **64**, 643–653, 2004.

[5] Gupta, S.N., Gupta, B.C. and Singh, S. *Estimation of sensitivity level of personnel interview survey questions*, Journal of Statistical Planning and Inference **100**, 239–247, 2002.

[6] Gupta, S., Shabbir, J. and Sehra, S. *Mean and sensitivity estimation in optional randomized response models*, Journal of Statistical Planning and Inference **140** (10), 2870–2874, 2010.

[7] Gupta, S., Shabbir, J., Sousa, R. and Real, P.C. *Estimation of the mean of a sensitive variable in the presence of auxiliary information*, Communications in Statistics-Theory and Methods **41**, 1–12, 2012.

[8] Kalucha, G., Gupta, S. and Dass, B.K. *Ratio estimation of Finite population Mean Using Optional Randomized Response Models*, to appear in the Journal of Statistical Theory and Practice **9** (3), 633–645, 2015.

[9] Mukhopadhyay, P. *Theory and Methods of Survey Sampling*, New Delhi: Prentice Hall of India, 1998.

[10] Singh, H.P. and Tarray, T.A. *A stratified Mangat and Singh's optional randomized response model using proportional and optimal allocation*, Stataistica **73** (1), 61–79, 2014.

[11] Sousa, R., Shabbir, J., Corte-Real, P. and Gupta, S. *Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information*, Journal of Statistical Theory and Practice **4** (3), 495–507, 2010.

[12] Sukhatme, P.V. and Sukhatme, B.V., *Sampling Thoery of Surveys with Applications*, Iowa State University Press, 1970.

[13] Wu, J.-W., Tian, G.-L. and Tang, M.-L. *Two new models for survey sampling with sensitive characteristic: Design and analysis*, Metrika **67**, 251–263, 2008.

[14] Warner, S.L. *Randomized response: a survey technique for eliminating evasive answer bias*, Journal of the American Statistical Association **60**, 63–69, 1965.