REVSTAT – Statistical Journal
Volume 18, Number 2, April 2020, 165–176

VARIANCE ESTIMATION USING RANDOMIZED RESPONSE TECHNIQUE

Authors: Sat Gupta
– Department of Mathematics and Statistics, University of North Carolina at Greensboro, North Carolina, USA
sngupta@uncg.edu

Badr Aloraini
– Department of Mathematics and Statistics, University of North Carolina at Greensboro, North Carolina, USA
boalorai@uncg.edu

Muhammad Nouman Qureshi
– National College of Business Administration and Economics, Lahore, Pakistan
nqureshi633@gmail.com

Sadia Khalil
– Department of Statistics, Lahore College for Women University, Lahore, Pakistan
sadia_khalil@hotmail.com

Received: January 2019 Revised: August 2019 Accepted: September 2019

Abstract:

- Variance estimation is a well-studied topic in survey sampling but not much work has been done in this area in the context of Randomized Response Technique (RRT) models. We propose here some variance estimators for sensitive variables using auxiliary information. We examine the performance of the proposed estimators through a simulation study and through a numerical example.

Key-Words:

- *auxiliary information; Mean Square Error; Randomized Response Technique; respondent privacy; variance estimation.*

AMS Subject Classification:

- 62D05.

1. INTRODUCTION

When conducting surveys, it is sometimes difficult to make a direct observation on the variable of interest. This is more so in the case where the research involves a topic that is a taboo in nature. In surveys on such topics, some of the respondents might give false responses. To offer a solution to this, a Randomized Response Technique (RRT) was developed by Warner [7]. The technique allows respondents to provide a response while maintaining their privacy.

The problem of mean and variance estimation is a topic that has been explored very well by researchers, although less so the problem of variance estimation. This is particularly the case in the context of RRT models. This is the main focus of this study where we examine variance estimation of a sensitive study variable using a highly correlated but non-sensitive auxiliary variable. According to Collins *et al.* [1], the auxiliary variables when combined with the main study variable help to achieve more efficient estimators.

In this paper, three variance estimators have been proposed under RRT using one auxiliary variable and two scrambling variables. In Section 2, some of the variance estimators in literature are reviewed. In Section 3, we propose a new class of variance estimators under RRT and derive their Bias as well as their MSE. We provide a comparison of the proposed estimators in Section 4. A numerical study is conducted in Section 5 based on real data. Some concluding remarks are given in Section 6.

2. ESTIMATORS IN LITERATURE

Let a simple random sample of size n be extracted without replacement from a finite population $U = \{U_1, U_2, ..., U_N\}$. Let Y be a sensitive variable of interest and X be a positively correlated auxiliary variable. Let (x_i, y_i) be the observed (X, Y) values for the i-th population unit U_i. Let $(\bar{x}, \bar{y})$ and $(\bar{X}, \bar{Y})$ be the sample and population means, and (s_x^2, s_y^2) and (σ_x^2, σ_y^2) be the sample and population variances respectively. Let

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\,, \qquad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\,,$$

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2\,, \qquad \sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2\,,$$

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N}X_i\,, \quad \bar{Y} = \frac{1}{N}\sum_{i=1}^{N}Y_i\,, \quad \bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i\,, \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N}y_i\,.$$

An unbiased estimator for the finite population variance is the sample variance given by:

$$t_0 = s_y^2\,.$$

Up to the first degree of approximation, its variance is given by

$$V(t_0) = \theta\sigma_y^4(\lambda_{40} - 1),$$

where

$$\lambda_{rs} = \frac{\mu_{rs}}{\mu_{20}^{\frac{r}{2}}\,\mu_{02}^{\frac{s}{2}}}, \qquad \mu_{rs} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^r (X_i - \bar{X})^s, \qquad \text{and} \qquad \theta = \frac{1}{n}.$$

Also 'r' and 's' are non-negative integers, μ_{20} and μ_{02} are the second order moments and λ_{rs} is the moment ratio.

Isaki [4] proposed the following ratio estimator of population variance using auxiliary information:

$$t_1 = s_y^2\left(\frac{\sigma_x^2}{s_x^2}\right).$$

The expressions for Bias and Mean Square Error (MSE) of the estimator, up to the first order of approximation, are given by

$$\mathrm{B}(t_1) = \theta\sigma_y^2(\lambda_{04} - 1)\,[1 - f_{04}]$$

and

$$\mathrm{MSE}(t_1) = \theta\sigma_y^4(\lambda_{40} - 1) + (\lambda_{04} - 1)\,[1 - 2f_{04}],$$

where

$$f_{04} = \frac{(\lambda_{22} - 1)}{(\lambda_{04} - 1)}.$$

The regression estimator of population variance was also proposed by Isaki [4] as

$$t_2 = s_y^2 + \alpha(\sigma_x^2 - s_x^2), \qquad \text{where} \quad \alpha = \left(\frac{\sigma_y^2}{\sigma_x^2}\right) f_{04}.$$

The MSE of t_2 is given by

$$\mathrm{MSE}(t_2) = \theta\sigma_y^4(\lambda_{40} - 1)(1 - p^2), \qquad \text{where} \quad p = (\lambda_{22} - 1)/\sqrt{(\lambda_{40} - 1)(\lambda_{04} - 1)}.$$

3. PROPOSED ESTIMATORS

Since Y is sensitive in nature, and hence subject to social desirability bias, we observe only a scrambled version of Y as given by Diana and Perri [2]. This is given by $Z = TY + S$, where T and S are scrambling variables. We also assume that Y, T and S are mutually uncorrelated. We also assume $E(S) = 0$ and $E(T) = 1$.

To obtain the Bias and MSE expressions for the proposed estimators, we define the following error terms:

$$s_z^2 = \sigma_z^2(1 + \delta_z) \qquad \text{and} \qquad \bar{z} = \bar{Z}(1 + e_z),$$

where

$$\delta_z = \frac{s_z^2 - \sigma_z^2}{\sigma_z^2} \qquad \text{and} \qquad e_z = \frac{\bar{z} - \bar{Z}}{\bar{Z}}$$

such that

$$E(\delta_z) = E(e_z) = 0\,, \quad E(\delta_z^2) = \theta(\lambda_{40} - 1)\,, \quad \text{and} \quad E(e_z^2) = \theta C_z^2\,; \quad \text{and} \quad E(\delta_z e_z) = \theta \lambda_{30} C_z$$

where

$$C_z^2 = C_y^2 \sigma_T^2 + \left(\frac{\sigma_S^2}{\bar{Y}^2}\right).$$

We now propose several population variance estimators under RRT.

3.1. A basic variance estimator under RRT

Based on the RRT model $Z = TY + S$, we have σ_z^2 as

$$\begin{aligned} \sigma_z^2 &= \sigma_{TY+S}^2 = \sigma_{TY}^2 + \sigma_S^2 \\ &= \left(\sigma_T^2 * \sigma_Y^2 + \sigma_T^2 * \big(E[Y]\big)^2 + \big(E[T]\big)^2 * \sigma_Y^2\right) + \sigma_S^2 \\ &= \left(\sigma_T^2 * \sigma_Y^2 + \sigma_T^2 * (\mu_Y)^2 + \sigma_Y^2\right) + \sigma_S^2 \\ &= \sigma_T^2 * \sigma_Y^2 + \sigma_T^2 * \mu_Y^2 + \sigma_Y^2 + \sigma_S^2\,. \end{aligned}$$

Rearranging, we get

$$\sigma_y^2 = \frac{\sigma_z^2 - \sigma_S^2 - (\sigma_T^2 * \bar{Z}^2)}{\sigma_T^2 + 1}\,.$$

Estimating σ_z^2 by its unbiased estimator s_z^2, we have our first proposed estimator given by

$$t_0(R) = \frac{s_z^2 - \sigma_S^2 - \sigma_T^2 * \bar{z}^2}{\sigma_T^2 + 1}\,. \tag{3.1}$$

Rewriting (3.1), we have

$$t_0(R) = \frac{\sigma_z^2(1+\delta_z) - \sigma_S^2 - \sigma_T^2\big[\bar{Z}(1+e_z)\big]^2}{\sigma_T^2 + 1}\,.$$

Subtracting σ_y^2 on both sides, we obtain

$$\big(t_0(R) - \sigma_y^2\big) = \frac{\sigma_z^2 \delta_z - 2\,\sigma_T^2 \bar{Z}^2 e_z - \sigma_T^2 \bar{Z}^2 e_z^2}{\sigma_T^2 + 1}\,. \tag{3.2}$$

By taking the expectation on both sides of (3.2), the Bias of $t_0(R)$ is obtained as

$$\text{Bias}\big(t_0(R)\big) = -\theta\left(\frac{\sigma_T^2 \bar{Z}^2}{\sigma_T^2 + 1}\right) C_z^2\,.$$

By squaring both sides of (3.2) and using the first order approximation, the MSE is obtained as

$$\text{MSE}\big(t_0(R)\big) = \theta\left(\frac{1}{(\sigma_T^2+1)^2}\right)\left(\sigma_z^4(\lambda_{40}-1) + 4\,\sigma_T^4 \bar{Z}^4 C_z^2 - 4\,\sigma_z^2 \sigma_T^2 \bar{Z}^2 \lambda_{30} C_z\right). \tag{3.3}$$

3.2. The ratio estimator under RRT

Isaki [4] proposed the classical ratio estimator $t_1 = s_y^2\left(\frac{\sigma_x^2}{s_x^2}\right)$. The RRT version of t_1 is

$$t_1(R) \;=\; \frac{s_z^2 - \sigma_S^2 - \sigma_T^2 * \bar{z}^2}{\sigma_T^2 + 1} * \left(\frac{\sigma_x^2}{s_x^2}\right). \tag{3.4}$$

To obtain the Bias and MSE, we define the following error terms:

$$s_x^2 = \sigma_x^2(1+\delta_x), \qquad \text{where} \quad \delta_x = \frac{s_x^2 - \sigma_x^2}{\sigma_x^2},$$

such that

$$E(\delta_x) = 0, \qquad E(\delta_x^2) = \theta(\lambda_{04} - 1) \qquad \text{and} \qquad E(\delta_x e_z) = \theta\lambda_{12} C_z .$$

Rewriting (3.4), we have

$$t_1(R) \;=\; \frac{\sigma_z^2 - \sigma_S^2 - \sigma_T^2 \bar{Z}^2}{\sigma_T^2 + 1} + \frac{2\,\sigma_T^2 \bar{Z}^2 e_z \delta_x - \sigma_z^2 \delta_z \delta_x - \sigma_T^2 \bar{Z}^2 e_z^2}{\sigma_T^2 + 1}.$$

Subtracting σ_y^2 and taking the expectation on both sides, the Bias of $t_1(R)$ is obtained as

$$\text{Bias}\big(t_1(R)\big) \;=\; \theta\left(\frac{2\,\sigma_T^2 \bar{Z}^2 \lambda_{12} C_z - \sigma_z^2(\lambda_{22}-1) - \sigma_T^2 \bar{Z}^2 C_z^2}{\sigma_T^2 + 1}\right). \tag{3.5}$$

For MSE, we have

$$t_1(R) \;=\; \frac{\sigma_z^2 + \sigma_z^2\delta_z - \sigma_S^2 - \sigma_T^2 \bar{Z}^2 - 2\,\sigma_T^2 \bar{Z}^2 e_z - \sigma_T^2 \bar{Z}^2 e_z^2}{\sigma_T^2 + 1}$$
$$\frac{-\sigma_z^2\delta_x - \sigma_z^2\delta_z\delta_x + \sigma_S^2\delta_x + \sigma_T^2 \bar{Z}^2\delta_x + 2\,\sigma_T^2 \bar{Z}^2 e_z\delta_x + \sigma_T^2 \bar{Z}^2 e_z^2\delta_x}{\sigma_T^2 + 1}.$$

Simplifying and ignoring second and higher order terms,

$$t_1(R) \;=\; \frac{\sigma_z^2 - \sigma_S^2 W - \sigma_T^2 \bar{Z}^2}{\sigma_T^2 + 1} + \frac{\sigma_z^2\delta_z - 2\,\sigma_T^2 \bar{Z}^2 e_z - \sigma_z^2\delta_x + \sigma_S^2\delta_x + \sigma_T^2 \bar{Z}^2\delta_x}{\sigma_T^2 + 1}.$$

Squaring and taking the expectation on both sides, we have

$$\text{MSE}\big(t_1(R)\big) \;=\; E\left(\frac{\sigma_z^2\delta_z}{\sigma_T^2 + 1} - \frac{2\,\sigma_T^2 \bar{Z}^2 e_z}{\sigma_T^2 + 1} - \sigma_y^2\delta_x\right)^2.$$

After some simplifications, the MSE of $t_1(R)$ is obtained as

$$\begin{aligned}\text{MSE}\big(t_1(R)\big) \;=\; \theta\frac{1}{(\sigma_T^2+1)^2}\Big[&\sigma_z^4(\lambda_{40}-1) - 2\,\sigma_z^2\sigma_y^2(\lambda_{22}-1)(\sigma_T^2+1) + \sigma_y^4(\lambda_{04}-1)(\sigma_T^2+1)^2 \\ &+\; 4C_z\Big(\sigma_T^4 \bar{Z}^4 C_z - \sigma_z^2\sigma_T^2 \bar{Z}^2\lambda_{30} + \sigma_T^2\sigma_y^2 \bar{Z}^2\lambda_{12}(\sigma_T^2+1)\Big)\Big].\end{aligned} \tag{3.6}$$

3.3. A generalized variance estimator under RRT

We now propose the following class of generalized population variance estimators:

$$t_p(R) = \left[\left(\left(\frac{s_z^2 - \sigma_S^2 - \sigma_T^2 * \bar{z}^2}{\sigma_T^2 + 1}\right) + (\sigma_x^2 - s_x^2)\right) * \left(\frac{(\alpha\sigma_x^2 + \beta)}{\omega(\alpha s_x^2 + \beta) + (1-\omega)(\alpha\sigma_x^2 + \beta)}\right)^g\right], \tag{3.7}$$

where g, α, β and ω are suitably chosen constants. We would choose $g = 1$ for positive correlation between Y and X, and -1 for negative correlation. α and β are known parameters associated with the auxiliary variable and ω is obtained from optimality consideration.

Using Taylor series approximation, we obtain the bias of the generalized estimator $t_p(R)$ as

$$\mathrm{Bias}\big(t_p(R)\big) = \frac{-\theta\,\sigma_T^2\bar{Z}^2}{\sigma_T^2 + 1}C_z^2 - (g\,\omega\,\psi_i)\,\theta\left(\frac{\sigma_z^2(\lambda_{22}-1) - 2\,\sigma_T^2\bar{Z}^2\lambda_{12}C_z}{\sigma_T^2 + 1} - \sigma_x^2(\lambda_{04}-1)\right), \tag{3.8}$$

where $\psi_i = \frac{\alpha\sigma_x^2}{\alpha\sigma_x^2 + \beta}$.

The mean square error is given by

$$\begin{aligned}\mathrm{MSE}\big(t_p(R)\big) = \theta\Bigg[&\left(\frac{\sigma_z^4(\lambda_{40}-1) + 4\,\sigma_T^4\bar{Z}^4C_z^2 - 4\,\sigma_z^2\sigma_T^2\bar{Z}^2\lambda_{30}C_z}{(\sigma_T^2+1)^2}\right)\\ &+ \Big((\sigma_x^2 + Q\,\sigma_y^2)^2\,(\lambda_{04}-1)\Big)\\ &- 2\left(\frac{\sigma_z^2(\lambda_{22}-1) - 2\,\sigma_T^2\bar{Z}^2\lambda_{12}C_z}{\sigma_T^2+1}\right)(\sigma_x^2 + Q\,\sigma_y^2)\Bigg],\end{aligned} \tag{3.9}$$

where $Q = g\,\omega\,\psi_i$.

Differentiate (3.9) w.r.t Q:

$$2\,\sigma_y^2(\sigma_x^2 + Q\,\sigma_y^2)\,(\lambda_{04}-1) = 2\,\sigma_y^2\left(\frac{\sigma_z^2(\lambda_{22}-1) - 2\,\sigma_T^2\bar{Z}^2\lambda_{12}C_z}{\sigma_T^2+1}\right),$$

$$Q_{\mathrm{opt}} = \frac{1}{\sigma_y^2}\left[\left(\frac{\sigma_z^2(\lambda_{22}-1) - 2\,\sigma_T^2\bar{Z}^2\lambda_{12}C_z}{\sigma_T^2+1}\right)\left(\frac{1}{(\lambda_{04}-1)}\right) - \sigma_x^2\right].$$

The MSE at this optimum value is given by

$$\begin{aligned}\mathrm{MSE}\big(t_p(R)\big)_{\mathrm{opt}} = \frac{\theta}{(\sigma_T^2+1)^2}\Bigg[&\Big(\sigma_z^4(\lambda_{40}-1) + 4\,\sigma_T^4\bar{Z}^4C_z^2 - 4\,\sigma_z^2\sigma_T^2\bar{Z}^2\lambda_{30}C_z\Big)\\ &- \frac{1}{(\lambda_{04}-1)}\Big(\sigma_z^2(\lambda_{22}-1) - 2\,\sigma_T^2\bar{Z}^2\lambda_{12}C_z\Big)^2\Bigg].\end{aligned} \tag{3.10}$$

4. SIMULATION STUDY

In this section, we use a simulation study to evaluate how efficient the generalized estimator $t_p(R)$ is as compared to both the basic estimator $t_0(R)$ and the ratio estimator $t_1(R)$. We first consider samples of size $N = 1000$ each from three bivariate normal populations determined by the following means and covariance matrices:

$$\text{Population I:} \quad \mu = \begin{bmatrix} 6 \\ 4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 4 & 1.6 \\ 1.6 & 1 \end{bmatrix}, \quad \rho_{yx} = 0.80\,;$$

$$\text{Population II:} \quad \mu = \begin{bmatrix} 6 \\ 4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 4 & 2.25 \\ 2.25 & 2 \end{bmatrix}, \quad \rho_{yx} = 0.80\,; \tag{4.1}$$

$$\text{Population III:} \quad \mu = \begin{bmatrix} 6 \\ 4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 4 & 1.2 \\ 1.2 & 1 \end{bmatrix}, \quad \rho_{yx} = 0.60\,.$$

These 1000 observations are treated as our finite populations. For the 1000 values generated from these distributions, the means, variances, covariances, and correlations are given by

$$\text{Population I:} \quad \mu_x = 6.029\,, \quad \mu_y = 4.007\,, \quad \sigma_x^2 = 3.8862\,, \quad \sigma_y^2 = 0.9450\,, \quad \sigma_{xy} = 1.5284\,, \quad \rho_{yx} = 0.7975389\,;$$

$$\text{Population II:} \quad \mu_x = 6.021\,, \quad \mu_y = 3.9836\,, \quad \sigma_x^2 = 3.9467\,, \quad \sigma_y^2 = 1.9998\,, \quad \sigma_{xy} = 2.2382\,, \quad \rho_{yx} = 0.7967094\,;$$

$$\text{Population III:} \quad \mu_x = 5.962\,, \quad \mu_y = 3.971\,, \quad \sigma_x^2 = 4.1149\,, \quad \sigma_y^2 = 0.9560\,, \quad \sigma_{xy} = 1.2442\,, \quad \rho_{yx} = 0.5927674\,.$$

For each population, we consider samples of sizes 200 and 500. The scrambling variables S and T are assumed to have normal distributions with $E(S) = 0$ and $E(T) = 1$. We have used different values for $\mathrm{Var}(S)$ and $\mathrm{Var}(T)$.

Before presenting the simulation results, we would like to note that in most studies, researchers have compared estimators only with respect to the *Percent Relative Efficiency* which is defined as

$$\mathrm{PRE} = \frac{\mathrm{MSE}\big(t_0(R)\big)}{\mathrm{MSE}\big(t_i(R)\big)} \times 100\,, \qquad \text{where} \quad i = 0, 1 \text{ and } p\,.$$

However, for estimators based on RRT methodology, one needs to also consider the *Privacy Protection* offered by the RRT model. With that in mind, Gupta *et al.* [3] introduced a unified measure of estimator quality (δ) given by

$$\delta = \frac{\text{Theoretical MSE}}{\Delta_{DP}}\,, \qquad \text{where} \quad \Delta_{DP} = E(Z-Y)^2 = \sigma_T^2(\mu_y^2 + \sigma_y^2) + \sigma_s^2$$

is the privacy level for the model $Z = TY + S$, as per Yan *et al.* [8]. A smaller value of (δ) is to be preferred. Khalil *et al.* [6] used this unified measure to compare the performance of various mean estimators under RRT.

Table 1: Theoretical (**bold**) and empirical MSEs and PREs of the estimators for Population I with $\sigma_T^2 = 0.5$, $\sigma_y^2 = 1$ and $\rho_{yx} = 0.80$.

Var(S)	n	Estimator	Mean($\hat{\sigma}_y^2$)	MSE	PRE	δ
0.2	200	$t_0(R)$	1.018416	**0.4593715** 0.4629093	**100** 100	0.052801
		$t_1(R)$	0.9873038	**0.4166811** 0.4137594	**110.2453** 111.8788	0.04789438
		$t_p(R)$	0.9708478	**0.3685766** 0.3689481	**124.6339** 125.4673	0.04236513
	500	$t_0(R)$	1.021572	**0.1995375** 0.2100302	**100** 100	0.02293534
		$t_1(R)$	0.9846944	**0.1430092** 0.146612	**139.5277** 143.2558	0.01643784
		$t_p(R)$	0.9999683	**0.0946721** 0.0957580	**210.7669** 219.3343	0.01088185
0.5	200	$t_0(R)$	1.034554	**0.5512713** 0.5593184	**100** 100	0.06125237
		$t_1(R)$	0.9986482	**0.4943552** 0.5045654	**111.5131** 110.8515	0.05492836
		$t_p(R)$	0.9854447	**0.4320352** 0.4187965	**127.5987** 133.5537	0.04800391
	500	$t_0(R)$	1.023019	**0.2022691** 0.1991505	**100** 100	0.02247434
		$t_1(R)$	0.9816713	**0.1866725** 0.182478	**108.3550** 109.1367	0.02074139
		$t_p(R)$	1.00232	**0.1686173** 0.1685935	**119.9575** 118.1246	0.01873526
1	200	$t_0(R)$	1.032376	**0.6313128** 0.6288249	**100** 100	0.06645398
		$t_1(R)$	0.9967019	**0.5716984** 0.5582806	**110.4275** 112.6359	0.06017878
		$t_p(R)$	0.9682892	**0.494106** 0.5058227	**127.7686** 124.3172	0.05201116
	500	$t_0(R)$	1.040029	**0.2705931** 0.2652877	**100** 100	0.02848348
		$t_1(R)$	0.9968461	**0.212635** 0.2254085	**127.2570** 117.6919	0.02238263
		$t_p(R)$	0.9791635	**0.1965888** 0.204013	**137.6442** 130.0347	0.02069356

Tables 1, 2 and 3 show the values of the theoretical MSEs and empirical MSEs. The values from the table confirm that the basic estimator $t_0(R)$ and the ratio estimator $t_1(R)$ are less efficient as compared to the generalized estimator $t_p(R)$. Also, while comparing the generalized estimator $t_p(R)$ with the ratio estimator $t_1(R)$ and basic estimator $t_0(R)$, we note that as the variance of T or variance of S increase, the MSEs increase. This is expected since adding more noise makes the MSE increase. However, if we look at the unified measure (δ), we find that it does not always increase as variance of T or variance of S increase, or at least not to the same extent as does the MSE. For example, for the generalized estimator

$t_p(R)$, theoretical MSE for Population II, with sample size 500, is 0.09227229 for $\sigma_T^2 = 0.2$ but increases to 0.3790013 for $\sigma_T^2 = 1$. In contrast, the (δ) value decreases from 0.023659 to 0.020499. Admittedly, this is not a big drop in (δ) value but at least it is not going up. The important point here is that the 310% increase in MSE (from 0.09227229 to 0.3790013) is more than offset by the significant increase in privacy level in using $\sigma_T^2 = 1$ as compared to $\sigma_T^2 = 0.2$. In another example, for the generalized estimator $t_p(R)$, theoretical MSE for Population III, with sample size 500, is 0.1877209 for $\sigma_T^2 = 0.5$ but increases to 0.3634541 for $\sigma_T^2 = 1$. In contrast, the (δ) value decreases from 0.021453 to 0.021069.

Table 2: Theoretical (**bold**) and empirical MSEs and PREs of the estimators for Population II with $\sigma_s^2 = 0.5$, $\sigma_y^2 = 2$ and $\rho_{yx} = 0.80$.

Var(T)	n	**Estimator**	**Mean($\hat{\sigma}_y^2$)**	**MSE**	**PRE**	δ
0.2	200	$t_0(R)$	1.961504	**0.3353948** 0.3330506	**100** 100	0.085998
		$t_1(R)$	1.938223	**0.3086746** 0.310405	**108.6564** 107.2955	0.079147
		$t_p(R)$	1.97547	**0.2604031** 0.2696629	**128.7983** 123.5062	0.066770
	500	$t_0(R)$	1.984015	**0.1299197** 0.1273284	**100** 100	0.033312
		$t_1(R)$	1.999045	**0.1057879** 0.1067183	**122.8114** 119.3126	0.027125
		$t_p(R)$	1.985764	**0.09227229** 0.09218931	**140.8003** 138.1162	0.023659
0.5	200	$t_0(R)$	1.997112	**0.8036853** 0.7958328	**100** 100	0.084651
		$t_1(R)$	1.988183	**0.7195406** 0.694571	**111.6942** 114.5790	0.075788
		$t_p(R)$	1.98627	**0.624445** 0.6421061	**128.7039** 123.9410	0.065772
	500	$t_0(R)$	1.991561	**0.2858802** 0.2751116	**100** 100	0.030111
		$t_1(R)$	1.982515	**0.2471334** 0.232594	**115.6784** 118.2797	0.026030
		$t_p(R)$	1.968053	**0.1816638** 0.1885275	**157.3677** 145.9265	0.019134
1	200	$t_0(R)$	1.981875	**1.170947** 1.167372	**100** 100	0.063335
		$t_1(R)$	2.002721	**1.014171** 0.5582806	**115.4585** 112.8290	0.054855
		$t_p(R)$	1.988997	**0.955732** 0.969496	**122.5183** 120.4101	0.051694
	500	$t_0(R)$	1.979819	**0.5567679** 0.531363	**100** 100	0.030114
		$t_1(R)$	1.998328	**0.4837988** 0.4790216	**115.0825** 110.9267	0.026168
		$t_p(R)$	1.971607	**0.3790013** 0.3843118	**146.9039** 138.2635	0.020499

Table 3: Theoretical (**bold**) and empirical MSEs and PREs of the estimators for Population III with $\sigma_s^2 = 0.25$, $\sigma_y^2 = 1$ and $\rho_{yx} = 0.60$.

Var(T)	n	Estimator	Mean($\widehat{\sigma}_y^2$)	MSE	PRE	δ
0.2	200	$t_0(R)$	1.021512	**0.2249759** 0.223441	**100** 100	0.061637
		$t_1(R)$	1.037037	**0.1962207** 0.1958733	**114.6545** 114.0742	0.053759
		$t_p(R)$	0.979563	**0.1733191** 0.1752187	**129.8044** 127.5212	0.047484
	500	$t_0(R)$	0.99568	**0.09192312** 0.09384772	**100** 100	0.025184
		$t_1(R)$	1.035195	**0.08558669** 0.08575554	**107.4035** 109.4363	0.023448
		$t_p(R)$	0.995747	**0.06216159** 0.06279393	**147.8776** 149.4534	0.017030
0.5	200	$t_0(R)$	0.9830188	**0.6333537** 0.6304459	**100** 100	0.072383
		$t_1(R)$	1.039288	**0.5491218** 0.5699384	**115.3393** 110.6164	0.062756
		$t_p(R)$	0.971143	**0.4907475** 0.5044131	**129.0589** 124.9860	0.056085
	500	$t_0(R)$	0.9941702	**0.2469968** 0.2442127	**100** 100	0.028228
		$t_1(R)$	0.9846135	**0.2070803** 0.2115374	**119.2758** 115.4465	0.023666
		$t_p(R)$	0.9992722	**0.1877209** 0.1827657	**131.5766** 133.6206	0.021453
1	200	$t_0(R)$	0.9571123	**1.166476** 1.148805	**100** 100	0.067621
		$t_1(R)$	0.9954355	**1.092394** 1.087534	**106.7816** 105.6339	0.063327
		$t_p(R)$	0.9794743	**0.9463649** 0.9256485	**123.2585** 124.1081	0.054861
	500	$t_0(R)$	1.009706	**0.5152219** 0.4923866	**100** 100	0.029867
		$t_1(R)$	0.9918212	**0.4304643** 0.458314	**119.6898** 107.4343	0.024954
		$t_p(R)$	0.9856029	**0.3634541** 0.3569531	**141.7570** 137.9415	0.021069

5. APPLICATION

In this section, we use a real data to show the performance of the generalized estimator $t_p(R)$ in comparison to other estimators. For this data which can be obtained from James *et al.* [5], the population size is ($N = 777$). The study variable Y is the reported percent of alumni who donate. The auxiliary variable X is the student to faculty ratio. The scrambling variable S is taken to be a normal random variable with mean equal to zero and variance equal to 0.5. The scrambling variable T is taken to be a normal random variable with mean equal to 1 and variance equal to 0.2, 0.5, and 1.

Population Characteristics are given by

$$N = 777\,, \quad n = 200\,, \quad \mu_X = 14.08\,, \quad \mu_Y = 22.74\,,$$
$$\sigma_X = 3.95\,, \quad \sigma_Y = 12.39\,, \quad \sigma_{XY} = 19.7641\,, \quad \rho_{yx} = 0.40\,.$$

From the Table 4, it can be observed that the generalized estimator $t_p(R)$ performs better than the other estimators $t_0(R)$ and $t_1(R)$. Also, we can observe that the unified measure (δ) does not always increase as variance of T increases, or at least not to the same extent as does the MSE. For example, for the generalized estimator $t_p(R)$, theoretical MSE is 301.0716 for $\sigma_T^2 = 0.2$ but increases to 1196.559 for $\sigma_T^2 = 1$. In contrast, the (δ) value decreases from 2.23565474 to 1.78234135.

Table 4: Theoretical (**bold**) and empirical MSEs and PREs of the estimators.

n	Var(T)	Estimator	MSE	PRE	δ
500	0.2	$t_0(R)$	**519.1796** 490.2126	**100** 100	3.85525016
		$t_1(R)$	**435.4705** 437.4432	**119.2226** 112.0631	3.23365501
		$t_p(R)$	**301.0716** 297.7625	**172.4438** 164.6320	2.23565474
	0.5	$t_0(R)$	**896.6322** 888.2846	**100** 100	2.66917897
		$t_1(R)$	**643.4997** 620.5305	**139.3368** 143.1492	1.91563036
		$t_p(R)$	**596.1386** 570.0859	**150.4066** 155.8159	1.77464139
	1	$t_0(R)$	**1805.427** 1876.467	**100** 100	2.68928418
		$t_1(R)$	**1618.569** 1650.915	**111.5446** 113.6622	2.41094877
		$t_p(R)$	**1196.559** 1105.511	**150.8849** 169.7375	1.78234135

6. CONCLUSION

We propose here some variance estimators under RRT. These are the basic estimator $t_0(R)$, ratio estimator $t_1(R)$ and the generalized estimator $t_p(R)$. The simulation study reveals that the generalized estimator $t_p(R)$ is more efficient than the other estimators $t_0(R)$ and $t_1(R)$. We also examine the efficiency of the estimators relative to not just the MSE values, but also with respect to the unified measure of estimators quality (δ) and observe that while MSE always increases as the noise level increases, the (δ) value does not necessary follow this pattern. This highlights the significance of respondent under privacy.

ACKNOWLEDGMENTS

The authors would like to thank the two reviewers for their constructive comments which helped improve the presentation of this paper.

REFERENCES

[1] Collins, L.M.; Schafer, J.L.; *et al.* (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures, *Psychological Methods*, **6**(4), 330–351.

[2] Diana, G. and Perri, P.F. (2011). A class of estimators for quantitative sensitive data, *Statistical Papers*, **52**(3), 633–650.

[3] Gupta, S.; Samridhi, M.; Shabbir, J. and Khalil, S. (2018). A unified measure of respondent privacy and model efficiency in quantitative RRT models, *Journal of Statistical Theory and Practice*, **12**(3), 506–511.

[4] Isaki, C.T. (1983). Variance estimation using auxiliary information, *Journal of the American Statistical Association*, **78**(381), 117–123.

[5] James, G.; Witten, D.; Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*, Springer, New York.

[6] Khalil, S.; Zhang, Q. and Gupta, S. (2019). Mean estimation of sensitive variables under measurement errors using optional RRT models. To appear in *Communications in Statistics – Simulation and Computation.*

[7] Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association*, **60**(309), 63–69.

[8] Yan, Z.; Wang, J. and Lai, J. (2009). An efficiency and protection degree-based comparison among the quantitative randomized response strategies, *Communications in Statistics – Theory and Methods*, **38**, 404–408.