VFAST Transactions on Mathematics
http://vfast.org/journals/index.php/VTM@ 2024 ISSN(e): 2309-0022, ISSN(p): 2411-6343
Volume 12, Number 2, July-December, 2024 *pp: 17-31*

Efficient estimation of the population variance using a novel optional scrambling model

Muhammad Azeem[1*], **Abdul Salam**[1], **Ammara Nawaz Cheema**[2], **Abdul Basit**[1]

[1]Department of Statistics, University of Malakand, Khyber Pakhtunkhwa, Pakistan;
[2]Department of Mathematics, Air University, Islamabad, Pakistan

Keywords:*Efficiency, estimator, privacy protection, population variance, scrambled response, sensitive variable.*
Subject Classification:
2020 Mathematics Subject Classification: 62D05, 62F10.

Journal Info:
Submitted:
August 25, 2024
Accepted:
September 20, 2024
Published:
September 30, 2024

Abstract In the past few years, survey researchers have developed variance estimators of sensitive variables under randomized response techniques. The available variance estimators utilize linear scrambling models in which all of the survey participants are forced to scramble their responses and thus hide their true responses. In practice, some of the respondents may have no problem in reporting their true response to the researcher. The current study finds that using a true response option produces more efficient estimates of population variance compared to a linear scrambling model. Additionally, we also suggest a new variance estimator of a sensitive variable of interest and analyze its algebraic properties using an auxiliary variable. We also conduct a simulation study to show the improvement over the existing estimators of the population variance.

***Correspondence Author Email Address:**
azeemstats@uom.edu.pk
DOI: 10.21015/vtm.v12i2.1913

1 Introduction

In survey sampling, researchers seek information from the respondents on sensitive variables. A simple approach is to use a direct interviewing method; however, direct questioning generally results in high rates of non-response and/or false response. An alternative procedure is to use the randomized response survey technique where the survey participants are asked to scramble their responses. The purpose of the scrambling process is to protect the privacy of the respondents, thus motivating them to participate in the

survey. First introduced by Warner [26], the randomized response survey technique has attracted survey researchers over the past few decades. The study of Warner [27] suggested survey researcher to use an additive random variable in the interview process. Eichhorn and Hayre [8] suggested the use multiplicative random variable for privacy protection. The optional scrambling model by Gupta et al. [9] proved a key development in the history of randomized response models as it addressed the difficulties associated with the earlier models. The study of Azeem and Ali [3] provided a neutral comparison of the efficiency of optional scrambling models. Recently, Azeem [2] developed the use of an exponential random variable in randomized response models.

The use of supplementary variable in parameter estimation was suggested by Cochran [5] with the introduction of a ratio-type estimator. Das and Tripathi [6] used an ancillary variable to estimate the variance of the population of interest. Isaki [14] proposed a ratio-type variance estimator which was based on a single auxiliary variable positively related to the main variable. Singh et al. [24] suggested new efficient estimators for a finite population variance. Kadilar and Cingi [15] also utilized auxiliary information to improve the efficiency of variance estimators. Subramani and Kumarapandiyan [25] utilized the population median of the ancillary variable to suggest an improved variance estimator. Gupta et al. [12] used a linear scrambling model to propose a generalized estimator for the variance of the population. The results of the research of Gupta et al. [12] revealed that their suggested estimator improved the precision of the Isaki's [14] variance estimator. Saleem et al. [13] have recently presented a novel scrambling model along with suggesting a generalized variance estimator by using auxiliary variables.

Azeem et al. [4], Lovig et al. [18], and Azeem [1] developed simple and weighted metrics for evaluation of scrambling models. Other studies related to various types of scrambling procedures include Gupta et al. [10], Zhang et al. [29], Khalil et al. [16], Singh et al. [21], Singh and Singh [23], Singh et al. [22], Narjis and Shabbir [19], Kumar et al. [17] and Shabbir and Gupta [20].

Using a new randomized response technique, this study proposes a new improved variance estimator using an auxiliary variable. We show that our suggested variance estimator is more efficient than the existing variance estimators in situations in which the variable under study is sensitive. We also show that our proposed optional scrambling model provides more efficient estimates of variance than the linear scrambling model.

2 Notations

Suppose a finite population has N units U_1,U_2,U_3 ... U_N and from this population, let a simple random sample of size n is chosen. We denote the main sensitive variable by Y and let X be an auxiliary variable. Let (x_i, y_i) be the notation for the observed value of variable (X, Y) for the i_th population unit (i=1,2,..., N). Further, we use the symbols $(\bar{x}, \bar{y})$ and $(\bar{X}, \bar{Y})$, respectively, for the sample and population mean of variable X and Y. Moreover, let (s_X^2, s_Y^2) and be the notations for the sample and population variances of variable X and Y, respectively. The means and variances can be mathematically written as follows.

$$\sigma_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2,$$

$$\sigma_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2,$$

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

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

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

Define the moment ratio as:

$$\lambda_{TS} = \frac{\mu_{TS}}{\mu_{20}^{\frac{r}{2}}\mu_{02}^{\frac{s}{2}}},$$

where, $\mu_{TS} = \frac{\sum_{i=1}^{N-1}(y_i-\bar{Y})^r)(x_i-\bar{X})^s}{N-1}$, '$r$' and '$s$' are some integers greater than zero.

3 The Linear Combination Model (R1)

Let Z denotes the reported response, and let S be an additive scrambling/random variable, with, $E(S) = 0$, $E(S^2) = \sigma_S^2$, and $Var(S) = \sigma_S^2$. Before presenting our proposed estimator, we present some of the existing variance estimators using the linear combination model.

The linear combination model of Diana and Perri [7] can be expressed as:

$$Z = TY + S. \tag{1}$$

The measure of privacy protection associated with the above model can be derived as:

$$\Delta_{R1} = E(Z - Y)^2 = \sigma_T^2(\sigma_Y^2 + \mu_Y^2) + \sigma_S^2. \tag{2}$$

Using the linear combination model $Z = TY + S$, the variance of variable Z may be expressed as:

$$\sigma_Z^2 = \sigma_{TY}^2 + \sigma_S^2 = \sigma_T^2\sigma_Y^2 + \sigma_T^2\mu_Y^2 + \sigma_S^2,$$

or,

$$\sigma_Y^2 = \frac{\sigma_Z^2 - \sigma_S^2 - \sigma_T^2\bar{Z}^2}{\sigma_T^2 + 1}. \tag{3}$$

A simple variance estimator t_0 using the linear model may be obtained if we replace $\bar{Z}^2$ and σ_Z^2 by $\bar{z}^2$, s_z^2 , respectively. This yields the estimator:

$$t_0(R1) = \frac{s_z^2 - \sigma_S^2 - \sigma_T^2\bar{z}^2}{\sigma_T^2 + 1}. \tag{4}$$

The variance of the estimator t_0 can be derived as:

$$Var(t_0) = \theta\sigma_y^4(\lambda_{40} - 1), \tag{5}$$

where, $\theta = \frac{1}{n}$.
Isaki [9] developed a novel ratio-type estimator of the finite population variance, which is given as:

$$t_1 = s_y^2 \left(\frac{\sigma_X^2}{s_x^2} \right), \tag{6}$$

The Isaki's [9] variance estimator using the linear combination model may be written as:

$$t_1(R1) = \frac{s_z^2 - \sigma_S^2 - \sigma_T^2 \bar{z}^2}{\sigma_T^2 + 1} \left(\frac{\sigma_X^2}{s_x^2} \right). \tag{7}$$

The MSE of $t_1(R1)$, up to the first order of approximation, can be written in the form:

$$\begin{aligned} MSE(t_1(R1)) = & \frac{\theta}{(\sigma_T^2 + 1)^2} \left[\sigma_Z^4 (\lambda_{40} - 1) - 2\sigma_Z^2 \sigma_Y^2 (\sigma_T^2 + 1)(\lambda_{22} - 1) + \sigma_Y^4 (\sigma_T^2 + 1)^2 (\lambda_{04} - 1) \right] \\ & + \frac{\theta}{(\sigma_T^2 + 1)^2} \left[4C_Z (\sigma_T^4 \bar{Z}^4 C_Z - \sigma_Z^2 \sigma_T^2 \bar{Z}^2 \lambda_{30} + \sigma_T^2 \sigma_Z^2 \bar{Z}^2 \lambda_{12} (\sigma_T^2 + 1)) \right], \end{aligned} \tag{8}$$

where,

$$C_Z^2 = C_y^2 + \sigma_T^2 + \frac{\sigma_S^2}{\bar{Y}^2}.$$

Utilizing the linear scrambling model $Z = TY + S$, Gupta et al. [12] introduced a general estimator of the variance which is given as:

$$t_2(R1) = \left[\left(\left(\frac{s_z^2 - \sigma_S^2 - \sigma_T^2 \bar{z}^2}{\sigma_T^2 + 1} \right) + (\sigma_X^2 - s_X^2) \right) \left(\frac{\alpha \sigma_X^2 + \beta}{w(\alpha s_X^2 + \beta) + (1 - w)(\alpha \sigma_X^2 + \beta)} \right)^g \right]. \tag{9}$$

where g, w, α , and β denote some constants predetermined by the researcher.

The optimum mean square error of the Gupta et al. [12] variance estimator using the linear combination model (R1) can be obtained as:

$$\begin{aligned} MSE(t_2(R1))_{opt} = & \frac{\theta}{(\sigma_T^2 + 1)^2} \left[(\sigma_Z^4 (\lambda_{40} - 1) + 4\sigma_T^4 \bar{z}^4 C_Z^2 - 4\sigma_Z^2 \sigma_T^2 \bar{z}^2 \lambda_{30} C_Z) \right] \\ & - \frac{\theta}{(\sigma_T^2 + 1)^2} \left[\frac{1}{(\lambda_{04-1})} \left((\sigma_Z^2 (\lambda_{22} - 1) - 2\sigma_T^2 \bar{z}^2 \lambda_{12} C_Z \right)^2 \right]. \end{aligned} \tag{10}$$

4 Proposed Optional Scrambling Model (R2)

Using a scrambling variable S, the following optional randomized response model is proposed: written as:

$$Z = \begin{cases} (1 + \frac{\beta}{\alpha})Y + S & \text{with probability } \frac{\alpha}{\lambda(\alpha+\beta)} \\ \frac{Y}{2} - S & \text{with probability } \frac{\beta}{\lambda(\alpha+\beta)} \\ Y & \text{with probability } 1 - \frac{1}{\lambda}, \end{cases} \tag{11}$$

where α, β, and λ and are some pre-defined constants. The level of privacy under the general forced model may be derived as follows:

$$\Delta_P = E[Z - Y]^2.$$

This can be further simplified as:

$$\Delta_P = \frac{\alpha}{\lambda(\alpha+\beta)} E\left[\left(1+\frac{\beta}{2\alpha}\right)Y+S-Y\right]^2 + \frac{\beta}{\lambda(\alpha+\beta)} E\left(\frac{Y}{2}-S-Y\right)^2 + \left(1-\frac{1}{\lambda}\right)E[Y-Y]^2,$$

$$= \frac{\alpha}{\lambda(\alpha+\beta)} E\left[\frac{\beta}{2\alpha}Y+S\right]^2 + \frac{\beta}{\lambda(\alpha+\beta)} E\left[(-1)\left(\frac{Y}{2}+S\right)\right]^2,$$

$$= \frac{\alpha}{\lambda(\alpha+\beta)}\left[\frac{\beta^2}{4\alpha^2}(\sigma_Y^2+\mu_Y^2)+\sigma_S^2\right] + \frac{\beta}{\lambda(\alpha+\beta)}\left[\frac{1}{4}(\sigma_Y^2+\mu_Y^2)+\sigma_S^2\right],$$

$$= \left[\frac{\beta^2}{4\alpha\lambda(\alpha+\beta)} + \frac{\beta}{4\lambda(\alpha+\beta)}\right](\mu_Y^2+\sigma_Y^2) + \frac{1}{\lambda}\sigma_S^2,$$

$$\Delta_P = \frac{1}{\lambda}\left[\frac{\beta}{4\alpha}(\sigma_Y^2+\mu_Y^2)+\sigma_S^2\right]. \tag{12}$$

Using the proposed model, the variance of Z may be derived as:

$$Var(Z) = E(Z^2) - [E(Z)]^2. \tag{13}$$

In equation 13, $E[Z]$ may be simplified as:

$$E(Z) = \frac{\alpha}{\lambda(\alpha+\beta)} E\left[\left(1+\frac{\beta}{2\alpha}\right)Y+S\right] + \frac{\beta}{\lambda(\alpha+\beta)} E\left(\frac{Y}{2}-S\right) + \left(1-\frac{1}{\lambda}\right)E[Y],$$

$$E(Z) = \frac{\alpha}{\lambda(\alpha+\beta)}\left[\left(1+\frac{\beta}{2\alpha}\right)\mu_Y+0\right] + \frac{\beta}{\lambda(\alpha+\beta)}\left(\frac{\mu_Y}{2}-0\right) + \left(1-\frac{1}{\lambda}\right)\mu_Y.$$

Some mathematical simplification leads to:

$$E(Z) = \frac{2\alpha+\beta}{2\lambda(\alpha+\beta)}\mu_Y + \frac{\beta}{2\lambda(\alpha+\beta)}\mu_Y + \mu_Y - \frac{\mu_Y}{\lambda}.$$

Further simplification yields:

$$E(Z) = \mu_Y. \tag{14}$$

$E(Z)^2$ can be simplified as:

$$E(Z)^2 = \frac{\alpha}{\lambda(\alpha+\beta)} E\left[\left(1+\frac{\beta}{2\alpha}\right)Y+S\right]^2 + \frac{\beta}{\lambda(\alpha+\beta)} E\left(\frac{Y}{2}-S\right)^2 + \left(1-\frac{1}{\lambda}\right)E[Y^2],$$

$$= \frac{\alpha}{\lambda(\alpha+\beta)}\left[\left(1+\frac{\beta}{2\alpha}\right)^2(\sigma_Y^2+\mu_Y^2)+\sigma_S^2\right] + \frac{\beta}{\lambda(\alpha+\beta)}\left(\frac{1}{4}(\sigma_Y^2+\mu_Y^2)+\sigma_S^2\right) + \left(1-\frac{1}{\lambda}\right)(\sigma_Y^2+\mu_Y^2),$$

$$= \left[\frac{(2\alpha+\beta)^2+\alpha\beta}{4\alpha\lambda(\alpha+\beta)} + 1 - \frac{1}{\lambda}\right](\sigma_Y^2+\mu_Y^2) + \frac{1}{\lambda}\sigma_S^2.$$

Further simplification yields:

$$E(Z)^2 = \sigma_Y^2 + \mu_Y^2 + \frac{\beta}{4\alpha\lambda}(\sigma_Y^2 + \mu_Y^2) + \frac{1}{\lambda}\sigma_S^2. \tag{15}$$

Using equation 14 and equation 15 in equation 13 gives:

$$E(Z)^2 = \sigma_Z^2 = \left(1 + \frac{\beta}{4\alpha\lambda}\right)\sigma_Y^2 + \frac{\beta}{4\alpha\lambda}\bar{Z}^2 + \frac{1}{\lambda}\sigma_S^2. \tag{16}$$

Solving for σ_Y^2 yields:

$$\sigma_Y^2 = \frac{4\alpha\lambda}{4\alpha\lambda + \beta}\left(\sigma_Z^2 - \frac{1}{\lambda}\sigma_S^2 - \frac{\beta}{4\alpha\lambda}\bar{Z}^2\right). \tag{17}$$

The sample variance estimator may be expressed by replacing σ_Z^2 and $\bar{Z}^2$ by s_Z^2 and, $\bar{z}^2$respectively. This gives the following variance estimator:

$$t_0(R2) = \frac{4\alpha\lambda}{4\alpha\lambda + \beta}\left(s_Z^2 - \frac{1}{\lambda}\sigma_S^2 - \frac{\beta}{4\alpha\lambda}\bar{z}^2\right). \tag{18}$$

5 Proposed Estimator and its Properties

Motivated by Zaman and Bulut [28], we suggest the following variance estimator:

$$t_p = \left(s_Y^2 + \beta(\sigma_X^2 - s_X^2)\right) exp\left(\alpha\frac{\sigma_X^2 - s_X^2}{\sigma_X^2 + s_X^2}\right), \tag{19}$$

where α is a constant.

Under the linear combination model (R1), the proposed variance estimator may be written in the form:

$$t_p(R1) = \left(\frac{s_z^2 - \sigma_S^2 - \sigma_T^2\bar{z}^2}{\sigma_T^2 + 1} + \beta(\sigma_X^2 - s_x^2)\right) exp\left(\alpha\frac{\sigma_X^2 - s_x^2}{\sigma_X^2 + s_x^2}\right), \tag{20}$$

$$t_p(R2) = \left(\frac{4\alpha\lambda}{4\alpha\lambda + \beta}\left(s_Z^2 - \frac{1}{\lambda}\sigma_S^2 - \frac{\beta}{4\alpha\lambda}\bar{z}^2\right) + \beta(\sigma_X^2 - s_x^2)\right) exp\left(\alpha\frac{\sigma_X^2 - s_x^2}{\sigma_X^2 + s_x^2}\right). \tag{21}$$

Theorem 1. *Under the linear combination model (R1), the mean squared error of the suggested estimator can be written in the form:*

$$\begin{aligned} MSE(t_p(R1)) \approx \frac{\theta}{(\sigma_T^2 + 1)^2}[&(\sigma_z^4(\lambda_{40} - 1) + A^2C_Z^2 + B^2(\lambda_{04} - 1) \\ &+ 2AB\lambda_{12}C_Z - 2A\sigma_Z^2\lambda_{30}C_Z - 2B\sigma_Z^2(\lambda_{22} - 1)], \end{aligned} \tag{22}$$

where,

$$A = 2\sigma_T^2\bar{Z}^2, \tag{23}$$

and

$$B = \frac{\alpha}{2}(\sigma_Z^2 - \sigma_S^2 - \sigma_T^2\bar{Z}^2). \tag{24}$$

Proof. To derive the mean square error, let

$$s_Z^2 = \sigma_Z^2(1 + d_Z), s_X^2 = \sigma_X^2(1 + d_X), \bar{z} = \bar{Z}(1 + e_Z),$$

where

$$d_Z = \frac{s_Z^2 - \sigma_Z^2}{\sigma_Z^2}, d_X = \frac{s_X^2 - \sigma_X^2}{\sigma_X^2}, e_Z = \frac{\bar{z} - \bar{Z}}{\bar{Z}},$$

such that

$$E(d_Z) = E(d_X) = E(e_Z) = 0, E(d_Z)^2 = \theta(\lambda_{40} - 1), E(d_x)^2 = \theta(\lambda_{04} - 1), E(e_Z)^2 = \theta C_Z^2,$$

$$E(d_Z, d_X) = \theta(\lambda_{22} - 1), E(d_Z, e_Z) = \theta\lambda_{30}C_Z, E(d_X, e_Z) = \theta\lambda_{12}C_Z.$$

Using the above notations, our suggested variance estimator may be written as:

$$t_p(R1) = \left(\frac{\sigma_Z^2(1 + d_Z) - \sigma_S^2 - \sigma_T^2\bar{z}^2(1 + e_Z)^2}{\sigma_T^2 + 1} + \beta(\sigma_X^2 - \sigma_X^2(1 + d_X))\right) exp\left(\alpha\frac{\sigma_X^2 - \sigma_X^2(1 + d_X)}{\sigma_X^2 + \sigma_X^2(1 + d_X)}\right). \quad (25)$$

Simplifying and ignoring terms of higher order, equation 25 becomes:

$$t_p(R1) - \sigma_Y^2 \approx \frac{1}{\sigma_T^2 + 1}[\sigma_Z^2 d_Z - Ae_Z - Bd_X]. \quad (26)$$

Taking square on equation 26 and taking expectation on both sides leads to:

$$E[t_p(R1) - \sigma_Y^2]^2 \approx \frac{1}{(\sigma_T^2 + 1)^2}E[\sigma_Z^4 d_Z^2 + A^2 e_Z^2 + B^2 d_X^2 - 2Ae_Z\sigma_Z^2 d_Z + 2ABd_X e_Z - 2B\sigma_Z^2 d_Z d_X].$$

Further simplification yields:

$$MSE(t_p(R2)) \approx \frac{\theta}{(\sigma_T^2 + 1)^2}\left[(\sigma_z^4(\lambda_{40} - 1) + A^2 C_Z^2 + B^2(\lambda_{04} - 1)\right] + \frac{\theta}{(\sigma_T^2 + 1)^2}\left[2AB\lambda_{12}C_Z - 2A\sigma_Z^2\lambda_{30}C_Z - 2B\sigma_Z^2(\lambda_{22} - 1)\right].$$

□

Theorem 2. *Under the proposed model (R2), the mean squared error of the proposed estimator can be written in the form:*

$$MSE(t_p(R2)) \approx \frac{16\alpha^2\lambda^2\theta}{(4\alpha\lambda + \beta)^2}\left[(\sigma_Z^2(\lambda_{40} - 1) + D^2 C_Z^2 + G^2(\lambda_{04} - 1)\right] + \frac{16\alpha^2\lambda^2\theta}{(4\alpha\lambda + \beta)^2}\left[2DG\lambda_{12}C_Z - 2D\sigma_Z^2\lambda_{30}C_Z - 2G\sigma_Z^2(\lambda_{22} - 1)\right], \quad (27)$$

where,

$$D = \frac{\beta}{2\alpha\lambda}\bar{Z}^2, \quad (28)$$

and

$$G = \beta\sigma_X^2 - \frac{\alpha}{2}\sigma_Z^2 - \frac{\alpha}{2}\sigma_S^2 - \frac{\beta}{8\lambda}\bar{Z}^2. \quad (29)$$

Proof. Using these notations defined in Theorem 1, the proposed estimator may be expressed as:

$$t_p(R2) = \left(\frac{4\alpha\lambda}{4\alpha\lambda + \beta}\left(\sigma_Z^2(1 + d_Z) - \frac{1}{\lambda}\sigma_S^2 - \frac{\beta}{4\alpha\lambda}\bar{Z}^2(1 + e_Z)^2\right) + \beta(\sigma_X^2 - \sigma_X^2)(1 + d_X)\right) exp\left(\alpha\frac{\sigma_X^2 - \sigma_X^2(1 + d_X)}{\sigma_X^2 + \sigma_X^2(1 + d_X)}\right). \quad (30)$$

Simplifying and ignoring terms of order terms, equation 30 reduces to the form:

$$t_p(R2) - \sigma_Y^2 \approx \frac{4\alpha\lambda}{4\alpha\lambda + \beta}[\sigma_Z^2 d_Z - De_Z - Gd_X], \tag{31}$$

where D and G have been defined in equation 28 and equation 29, respectively.
Squaring equation 31 and applying expectation on both sides leads to:

$$E[t_p(R2) - \sigma_Y^2]^2 \approx \frac{16\alpha^2\lambda^2}{(4\alpha\lambda + \beta)^2} E[\sigma_Z^2 d_Z - De_Z - Gd_X]^2.$$

Further simplification yields:

$$\begin{aligned} MSE(t_p(R2)) \approx & \frac{16\alpha^2\lambda^2\theta}{(4\alpha\lambda + \beta)^2}\left[(\sigma_z^4(\lambda_{40} - 1) + D^2C_Z^2 + G^2(\lambda_{04} - 1)\right] \\ & + \frac{16\alpha^2\lambda^2\theta}{(4\alpha\lambda + \beta)^2}\left[2DG\lambda_{12}C_Z - 2D\sigma_Z^2\lambda_{30}C_Z - 2G\sigma_Z^2(\lambda_{22} - 1)\right]. \end{aligned}$$

□

Remark 1: The minimum values of D and G can be obtained by differentiating $MSE(t_p(R2))$ with respect to D and G to get the following equations.

$$DC_Z^2 = \sigma_Z^2\lambda_{30}C_Z - G\lambda_{12}C_Z, \tag{32}$$

and

$$G(\lambda_{04} - 1) = -D\lambda_{12}C_Z - \sigma_z^2(\lambda_{22} - 1). \tag{33}$$

Solving equation 32 and equation 33 gives the minimum values of D and G, expressed as follows:

$$D_{opt} = \sigma_Z^2\left(\frac{\lambda_{12}\lambda_{30} + \lambda_{22} - 1}{\lambda_{12}^2 - \lambda_{04} + 1}\right), \tag{34}$$

$$G_{opt} = \frac{\sigma_Z^2}{C_Z}\left[\frac{\lambda_{30}(\lambda_{12}^2 - \lambda_{04} + 1) - \lambda_{12}(\lambda_{12}\lambda_{30} + \lambda_{22} - 1)}{\lambda_{12}^2 - \lambda_{04} + 1}\right]. \tag{35}$$

Using these optimum values of D and G in equation 36 will yield the optimum value of the mean square error under the proposed model.
Remark 2: Using the notations introduced in Theorem 1, the bias of the proposed estimator using the proposed optional model can be easily obtained as:

$$Bias(t_p(R2)) \approx \frac{-4\alpha\lambda\theta}{4\alpha\lambda + \beta}\left[\frac{\beta\bar{Z}^2}{4\alpha\lambda}C_Z^2 + \frac{\alpha}{2}\sigma_Z^2(\lambda_{22} - 1) + \frac{\beta\bar{Z}^2}{4\lambda}\lambda_{12}C_Z + \frac{\alpha\beta}{2}\sigma_X^2(\lambda_{04} - 1)\right].$$

6 Comparison of Estimators

Under the linear combination model (R1), the proposed variance estimator is more efficient than the Isaki [14] variance estimator if:

$$MSE(t_p(R1) < MSE(t_1(R1), \tag{36}$$

Using equation 8 and equation 22 in 36 gives:

$$(A-4\sigma_T^4\bar{Z}^2)C_Z^2+[B^2-\sigma_Y^4(\sigma_T^2+1)^2](\lambda_{04}-1)< 2\left[\sigma_Z^2C_Z(A-2\sigma_T^2\bar{Z}^2)\lambda_{30}-c[AB-2\sigma_T^2\sigma_Y^2\bar{Z}^2(\sigma_T^2+1)\lambda_{12}\right] + 2\left[\sigma_Z^2[B-\sigma_Y^2(\sigma_T^2+1)](\lambda_{22}-1)\right]. \tag{37}$$

Under the proposed model (R2), the mean square error of the Isaki's [14] variance estimator can be derived as:

$$MSE(t_1(R2))\approx\frac{16\alpha^2\lambda^2\theta}{(4\alpha\lambda+\beta)^2}\left[(\sigma_z^4(\lambda_{40}-1)+H^2C_Z^2+K^2(\lambda_{04}-1)\right] + \frac{16\alpha^2\lambda^2\theta}{(4\alpha\lambda+\beta)^2}\left[2HK\lambda_{12}C_Z-2H\sigma_Z^2\lambda_{30}C_Z-2K\sigma_Z^2(\lambda_{22}-1)\right]. \tag{38}$$

Under the proposed model (R2), the proposed estimator performs more efficiently compared to the Isaki's [14] variance estimator if:

$$MSE(t_p(R2)<MSE(t_1(R2). \tag{39}$$

Using equation 27 and equation 38 in equation 39 gives:

$$\sigma_Z^2<\frac{2\lambda_{12}C_Z(HK-DG)-(G^2-K^2)(\lambda_{04}-1)}{(D^2-H^2)-2(D-H)C_Z\lambda_{30}-2(G-K)(\lambda_{22}-1)}. \tag{40}$$

Now comparing the proposed estimator under the linear combination model and under the proposed model, the efficiency condition can be obtained as:

$$MSE(t_p(R2)<MSE(t_p(R1). \tag{41}$$

Using equation 22 and equation 27 in equation 41, the above condition simplifies to:

$$\sigma_Z^2<\frac{C_Z^2(V^2A^2-U^2D^2)+(\lambda_{04}-1)(V^2B^2-U^2G^2)+2\lambda_{12}C_Z(V^2AB-U^2DG)}{(\lambda_{04}-1)(U^2-V^2)-2\lambda_{30}C_Z(U^2D-V^2A)-2(\lambda_{22}-1)(U^2G-V^2B)}, \tag{42}$$

where

$$U=\frac{4\alpha\lambda}{4\alpha\lambda+\beta},$$

and

$$V=\frac{1}{\sigma_T^2+1}.$$

A unified model-evaluation metric was suggested by Gupta et al. [11] as:

$$\delta=\frac{MSE}{\Delta}. \tag{43}$$

			α = 0.5, β = 0.5 α = 1.25, β = 3		α = 0.75, β = 0.75 α = 1.10, β = 2	
			Model		Model	
n	*W*	*Estimator*	*R*1	*R*2	*R*1	*R*2
50	0.4	t_0	0.129320	0.176658	0.0641126	-0.167691
		t_1	0.170575	0.214440	0.150889	-0.077342
		t_2	0.169661	0.213375	0.132252	-0.095333
		t_p	0.129805	0.176098	0.117649	-0.112014
50	0.8	t_0	0.129320	0.176658	0.064112	-0.167691
		t_1	0.170575	0.214440	0.150889	-0.077342
		t_2	0.371409	0.413466	0.255525	0.028949
		t_p	0.129805	0.176098	0.117649	-0.112014
150	0.4	t_0	-0.28759	-0.00994	0.081185	0.021145
		t_1	-0.27751	0.007468	0.076496	0.017953
		t_2	-0.27309	0.009355	0.062839	0.004353
		t_p	-0.28694	-0.00782	0.070097	0.010877
150	0.8	t_0	-0.28759	-0.00994	0.081185	0.021145
		t_1	-0.27751	0.007468	0.076496	0.017953
		t_2	-0.20923	0.081577	0.085986	0.028026
		t_p	-0.28694	-0.00782	0.070097	0.010877
250	0.4	t_0	0.021188	0.095099	0.099446	0.1510241
		t_1	0.032469	0.105845	0.101031	0.151734
		t_2	0.034243	0.107658	0.090682	0.141586
		t_p	0.023006	0.096773	0.0942025	0.145391
250	0.8	t_0	0.021188	0.095099	0.099446	0.151024
		t_1	0.032469	0.105845	0.101031	0.151734
		t_2	0.079445	0.152476	0.1072503	0.157763
		t_p	0.023006	0.096773	0.0942025	0.145391
400	0.4	t_0	0.019534	-0.034368	-0.022779	-0.019036
		t_1	0.030122	-0.023965	-0.023015	-0.018684
		t_2	0.034000	-0.019977	-0.027498	-0.023425
		t_p	0.023280	-0.030648	-0.025592	-0.021629
400	0.8	t_0	0.019534	-0.034368	-0.022779	-0.019036
		t_1	0.030122	-0.023965	-0.023015	-0.018684
		t_2	0.063879	0.009618	-0.018745	-0.014339
		t_p	0.023280	-0.030648	-0.025592	-0.021629

Table 1. Simulated Bias for μ_Y = 3, σ_Y^2 = 2, N = 5000, σ_T^2 = 0.25, and σ_S^2 = 5.

7 Simulation Study

For comparison of the performance of various estimators and randomized response models, we conducted a simulation study. We generated a population of N = 5000 units using R package. To maintain a correlation between the generated values for variable X and Y, we used a regression equation between Y and X. The results were averaged across 1000 iterations. The results of the simulated bias of the estimators under the linear combination model and the suggested optional scrambling model have been presented in Table 1. Likewise, the simulated mean squared errors and simulated values under each model can be observed in Table 2 and Table 3, respectively. Examining Table 2 and Table 3, the improvement in the proposed estimator over the already available estimators may clearly be observed. We also find that all of the variance estimators based on our proposed randomized response model are more efficient compared to the linear combination model. Further, we observe that the simulated values are also smaller under the proposed model.

			α = 0.5, β = 0.5 α = 1.25, β = 3		α = 0.75, β = 0.75 α = 1.10, β = 2	
			Model		Model	
n	*W*	*Estimator*	*R*1	*R*2	*R*1	*R*2
50	0.4	t_0	4.777006	3.803882	4.697353	3.670926
		t_1	4.892427	3.835163	5.014979	3.919348
		t_2	5.046275	4.008547	4.735513	3.740585
		t_p	4.569760	3.568900	4.573118	3.573921
50	0.8	t_0	4.777006	3.803882	4.697353	3.670926
		t_1	4.892427	3.835163	5.014979	3.919348
		t_2	7.234388	6.099096	5.722855	4.609517
		t_p	4.569760	3.568900	4.573118	3.573921
150	0.4	t_0	1.645065	1.237848	1.543035	1.169742
		t_1	1.597992	1.244827	1.512545	1.167188
		t_2	1.664439	1.339574	1.513488	1.174036
		t_p	1.557381	1.170619	1.464687	1.116583
150	0.8	t_0	1.645065	1.237848	1.543035	1.169742
		t_1	1.597992	1.244827	1.512545	1.167188
		t_2	2.116557	1.917187	1.760548	1.174036
		t_p	1.557381	1.170619	1.464687	1.116583
250	0.4	t_0	0.924019	0.725209	0.910408	0.744514
		t_1	0.908798	0.696082	0.951455	0.762017
		t_2	0.970557	0.970557	0.956403	0.772685
		t_p	0.876526	0.672867	0.907510	0.728406
250	0.8	t_0	0.924019	0.725209	0.910408	0.744514
		t_1	0.908798	0.696082	0.951455	0.762017
		t_2	1.301894	1.078280	1.131012	0.938165
		t_p	0.876526	0.672867	0.907510	0.728406
400	0.4	t_0	0.609564	0.449086	0.643009	0.450016
		t_1	0.594309	0.419566	0.630059	0.448316
		t_2	0.625153	0.452967	0.642336	0.459138
		t_p	0.576707	0.413399	0.623123	0.436776
400	0.8	t_0	0.609564	0.449086	0.643009	0.450016
		t_1	0.594309	0.419566	0.630059	0.448316
		t_2	0.805694	0.613119	0.713292	0.535991
		t_p	0.576707	0.413399	0.623123	0.436776

Table 2. Simulated values of MSEs for μ_Y = 3, σ_Y^2 = 2, N = 5000, σ_I^2 = 0.25, and σ_S^2 = 5.

8 Discussion and Conclusion

In the current study, we suggested a novel variance estimator of the sensitive variable under study, along with an improved optional scrambling technique. We analyzed the algebraic properties of our new proposed estimator using the new proposed scrambling model as well as under the existing linear combination model. We also derived the efficiency conditions for the proposed estimator under both models. Finally, we conducted a simulation study to assess the performance of our proposed variance estimator and our new proposed model.

Table 2 presents the simulated mean squared errors of different estimators of the population variance. The results were simulated across 1000 iterations with various sample sizes and different values of parameters. Table 2 clearly shows that our new suggested estimator is the most efficient of all estimators of the population variance. From Table 2, we also observe that our proposed optional scrambling model provides more efficient estimates of the variance than the estimates based on the linear model. It may also be observed that as the sample size n increases, the mean squared error of each variance estimator declines under each model.

For model-evaluation, the simulated values of have been provided in Table 3 under both the proposed

			α = 0.5, β = 0.5 α = 1.25, β = 3		α = 0.75, β = 0.75 α = 1.10, β = 2	
			Model		Model	
n	*W*	*Estimator*	*R*1	*R*2	*R*1	*R*2
50	0.4	t_0	0.571283	0.556446	0.572541	0.514092
		t_1	0.585086	0.561021	0.611255	0.548882
		t_2	0.603485	0.586385	0.577192	0.523847
		t_p	0.546498	0.522072	0.557399	0.500507
50	0.8	t_0	0.571283	0.556446	0.572541	0.514092
		t_1	0.585086	0.561021	0.611255	0.548882
		t_2	0.865162	0.892198	0.697536	0.645536
		t_p	0.546498	0.522072	0.557399	0.500507
150	0.4	t_0	0.199032	0.188236	0.181277	0.153979
		t_1	0.193337	0.189297	0.177695	0.153643
		t_2	0.201376	0.203705	0.177806	0.154544
		t_p	0.188424	0.178012	0.172072	0.146982
150	0.8	t_0	0.199032	0.188236	0.181277	0.153979
		t_1	0.193337	0.189297	0.177695	0.153643
		t_2	0.256077	0.291541	0.206830	0.186138
		t_p	0.188424	0.178012	0.172072	0.146982
250	0.4	t_0	0.111275	0.106514	0.111458	0.097900
		t_1	0.109443	0.102236	0.116483	0.100202
		t_2	0.116880	0.111332	0.117089	0.101605
		t_p	0.105556	0.098826	0.111103	0.095782
250	0.8	t_0	0.111275	0.106514	0.111458	0.097900
		t_1	0.109443	0.102236	0.116483	0.100202
		t_2	0.156782	0.158370	0.138465	0.123365
		t_p	0.105556	0.098826	0.111103	0.095782
400	0.4	t_0	0.073399	0.068327	0.078050	0.060748
		t_1	0.071562	0.063835	0.076478	0.060518
		t_2	0.097015	0.093284	0.077968	0.061979
		t_p	0.069442	0.062897	0.075636	0.058960
400	0.8	t_0	0.073399	0.068327	0.078050	0.060748
		t_1	0.071562	0.063835	0.076478	0.060518
		t_2	0.075276	0.068917	0.086581	0.072354
		t_p	0.069442	0.062897	0.075636	0.058960

Table 3. Simulated values of δ for μ_Y = 3, σ_Y^2 = 2, N = 5000, σ_T^2 = 0.25, and σ_S^2 = 5.

and the linear combination model. One may observe from Table 2 that the proposed variance estimator provides smaller values of. This makes our proposed model the most suitable model for real-world surveys in which the variable under consideration is sensitive. Based on the results of the study, we recommend the new proposed variance estimator and the proposed scrambling model for application in real-world sample surveys on sensitive variables.

The research work presented in this paper has some strengths and weaknesses. Optional randomized response models are superior to traditional models in that they offer the true response option, in addition to the scrambled response option. Unlike most of the available research studies which have proposed variance estimators under traditional randomized response models, we developed a novel optional randomized response model and proposed an efficient variance estimator under the new optional model. On the other negative side, the present study also has some limitations. The proposed estimator is only applicable in situations where response and non-response errors do not occur. Moreover, we evaluated the efficiency of our proposed variance estimator under simple random sampling which may not be suitable if the population units are heterogeneous.

9 Suggestions for Future Research

Measurement errors and non-response are two serious issues often encountered by survey statisticians. We recommend future researchers to work on analyzing the mathematical properties of our proposed variance estimator under the existence of measurement errors as well as non-response error.

Author Contributions

Muhammad Azeem: Conceptualization, Methodology, Data curation, Supervision, Writing- Original draft preparation.
Abdul Salam: Software, Data curation, Validation, Writing- Reviewing and Editing.
Ammara Nawaz Cheema: Conceptualization, Methodology, drafting, revision, Writing-revised version.
Abdul Abdul Basit: Software, simulation, drafting, proofreading.

Compliance with Ethical Standards

It is declared that the author doesn't have any conflict of interest. It is also declared that informed consent was obtained from all individual participants included in the study.

Funding Information

The authors have no funding to report.

Author Information

ORCID:
Muhammad Azeem: 0000-0002-6475-6072
Abdul Salam: 0009-0007-3056-8712
Ammara Nawaz Cheema: 0000-0002-3616-6024
Abdul Basit: 0009-0005-6116-7028

References

[1] Azeem, M. [2023*a*], 'Introducing a weighted measure of privacy and efficiency for comparison of quantitative randomized response models', *Pak. J. Statist* **39**(3), 377–385.

[2] Azeem, M. [2023*b*], 'Using the exponential function of scrambling variable in quantitative randomized response models', *Mathematical Methods in the Applied Sciences* **46**(13), 13882–13893.

[3] Azeem, M. and Ali, S. [2023], 'A neutral comparative analysis of additive, multiplicative, and mixed quantitative randomized response models', *Plos one* **18**(4), e0284995.

[4] Azeem, M., Salam, A., Albalawi, O. and Hussain, S. [2024], 'A new unified measure for evaluation of randomized response techniques', *Heliyon* **10**(16), e35852.

[5] Cochran, W. [1940], 'The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce', *The journal of agricultural science* **30**(2), 262–275.

[6] Das, A. K. [1978], 'Use of auxiliary information in estimating the finite population variance', *Sankhya, c* **40**, 139–148.

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

[8] Eichhorn, B. H. and Hayre, L. S. [1983], 'Scrambled randomized response methods for obtaining sensitive quantitative data', *Journal of Statistical Planning and inference* **7**(4), 307–316.

[9] Gupta, S., Gupta, B. and Singh, S. [2002], 'Estimation of sensitivity level of personal interview survey questions', *Journal of Statistical Planning and inference* **100**(2), 239–247.

[10] Gupta, S., Mehta, S., Shabbir, J. and Dass, B. [2013], 'Generalized scrambling in quantitative optional randomized response models', *Communications in Statistics-Theory and Methods* **42**(22), 4034–4042.

[11] Gupta, S., Mehta, S., 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**, 506–511.

[12] Gupta, S., Qureshi, M. N. and Khalil, S. [2020*a*], 'Variance estimation using randomized response technique', *REVSTAT-Statistical Journal* **18**(2), 165–176.

[13] Gupta, S., Qureshi, M. N. and Khalil, S. [2020*b*], 'Variance estimation using randomized response technique', *REVSTAT-Statistical Journal* **18**(2), 165–176.

[14] Isaki, C. T. [1983], 'Variance estimation using auxiliary information', *Journal of the American statistical association* **78**(381), 117–123.

[15] Kadılar, C. and Cingi, H. [2006], 'Improvement in variance estimation using auxiliary information', *Hacettepe Journal of mathematics and Statistics* **35**(1), 111–115.

[16] Khalil, S., Zhang, Q. and Gupta, S. [2021], 'Mean estimation of sensitive variables under measurement errors using optional rrt models', *Communications in Statistics-Simulation and Computation* **50**(5), 1417–1426.

[17] Kumar, S., Kour, S. P. and Singh, H. P. [2023], 'Applying orrt for the estimation of population variance of sensitive variable', *Communications in Statistics-Simulation and Computation* pp. 1–11.

[18] Lovig, M., Khalil, S., Rahman, S., Sapra, P. and Gupta, S. [2023], 'A mixture binary rrt model with a unified measure of privacy and efficiency', *Communications in statistics-simulation and computation* **52**(6), 2727–2737.

[19] Narjis, G. and Shabbir, J. [2023], 'An efficient new scrambled response model for estimating sensitive population mean in successive sampling', *Communications in Statistics-Simulation and Computation* **52**(11), 5327–5344.

[20] Shabbir, J. and Gupta, S. [2024], 'Estimation of sensitive trait proportion using kuk's randomized response model with auxiliary information', *Communications in Statistics-Theory and Methods* pp. 1–11.

[21] Singh, C., Singh, G. N. and Kim, J.-M. [2021], 'A randomized response model for sensitive attribute with privacy measure using poisson distribution', *Ain Shams Engineering Journal* **12**(4), 4051–4061.

[22] Singh, G. N., Singh, C. and Kumar, A. [2022], 'A modified randomized device for estimation of population mean of quantitative sensitive variable with measure of privacy protection', *Communications in Statistics-Simulation and Computation* **51**(4), 1867–1890.

[23] Singh, G. and Singh, C. [2022], 'Proficient randomized response model based on blank card strategy to estimate the sensitive parameter under negative binomial distribution', *Ain Shams Engineering Journal* **13**(5), 101611.

[24] Singh, S., Horn, S., Chowdhury, S. and Yu, F. [1999], 'Theory & methods: Calibration of the estimators of variance', *Australian & New Zealand Journal of Statistics* **41**(2), 199–212.

[25] Sumramani, J. and Kumarapandiyan, G. [2012], 'Variance estimation using median of the auxiliary variable', *International journal of Probability and Statistics* **1**(3), 36–40.

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

[27] Warner, S. L. [1971], 'The linear randomized response model', *Journal of the American Statistical Association* **66**(336), 884–888.

[28] Zaman, T. and Bulut, H. [2022], 'A new class of robust ratio estimators for finite population variance', *Scientia Iranica* pp. 1–25.

[29] Zhang, Q., Khalil, S. and Gupta, S. [2021], 'Mean estimation in the simultaneous presence of measurement errors and non-response using optional rrt models under stratified sampling', *Journal of statistical computation and Simulation* **91**(17), 3492–3504.