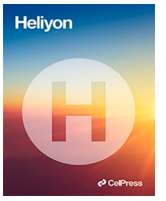
Heliyon 10 (2024) e27488

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

An efficient estimator of population variance of a sensitive variable with a new randomized response technique

Muhammad Azeem [a,*], Najma Salahuddin [b], Sundus Hussain [b], Musarrat Ijaz [c], Abdul Salam [a]

[a] Department of Statistics, University of Malakand, Khyber Pakhtunkhwa, Pakistan
[b] Department of Statistics, Shaheed Benazir Bhutto Women University, Peshawar, Pakistan
[c] Department of Statistics, Rawalpindi Women University, Rawalpindi, Pakistan

ARTICLE INFO

Keywords:
Auxiliary variable
Efficiency
Mean square error
Randomized response technique
Variance estimator

ABSTRACT

In sampling theory, a majority of the available estimators of population variance are designed for use with non-sensitive variables only. Such estimators cannot perform efficiently when the variable of interest is of sensitive nature, such as use of drugs, illegal income, abortion, cheating in examination, the amount of income tax payable, and the violation of rules by employees, etc. In the current literature, the shortage of research studies on variance estimators of a sensitive variable has created a big research gap and a room for improvement in the efficiency of such estimators. In this paper, a new randomized scrambling technique is proposed, along with a new estimator of population variance. The new estimator achieves improvement in efficiency over the available variance estimators. The proposed estimator is designed for use with simple random sampling and uses the information on an auxiliary variable. The improvement in efficiency is shown for different choices of constants. Besides efficiency, improvement in the unified measure of estimator quality is also achieved with the proposed estimator under the new randomized response model.

1. Introduction

Survey researchers are often faced with refusals and false responses in sample surveys on sensitive topics. Respondents often fail to provide truthful information on question related to sensitive issues such as abortion, use of drugs, illegal income, and cheating in examination. Introduced in 1965, the randomized response techniques can be a good alternative to the direct-questioning method when collecting information on sensitive variables. The Warner's [1] randomized response technique and its modified variants are designed to get sensitive information from the respondents yet protecting their privacy. Warner [2] introduced the use of a scrambling variable for the purpose of the respondents' privacy protection. A drawback of the Warner's [2] technique was that it forced every respondent to scramble his/her response even if the respondent perceived the question as non-sensitive. To alleviate this problem, Gupta et al. [3] introduced what is called the optional randomized response model. With optional randomization techniques, the respondents have the option to either provide the true response or go for reporting a scrambled response. The respondent's decision to report the true or a scrambled response depends on whether he/she perceives the question being asked as sensitive or not. Unlike the usual additive or multiplicative scrambling methods, a recent study of Azeem [4] introduced the idea of exponential scrambling

* Corresponding author.
E-mail address: azeemstats@uom.edu.pk (M. Azeem).

https://doi.org/10.1016/j.heliyon.2024.e27488
Received 7 August 2023; Received in revised form 21 February 2024; Accepted 29 February 2024
Available online 6 March 2024

Table 1
Notations for sample and population data.

Notation	Meaning
$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$	Population Mean of X
$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$	Population Mean of Y
$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	Sample Mean of X
$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	Sample Mean of Y
$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2$	Population Variance of X
$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2$	Population Variance of Y
$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$	Sample Variance of X
$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$	Sample Variance of Y
$\lambda_{rs} = \frac{\mu_{rs}}{\overline{\mu}_{20}^{\frac{r}{2}}\overline{\mu}_{02}^{\frac{s}{2}}}$	Moment ratio, where 'r' and 's' denote positive integers.
$\mu_{rs} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^r(x_i - \overline{X})^s$	Cross-product moment, where 'r' and 's' denote positive integers.

technique.

In sampling theory, Cochran [5] laid the foundation of using auxiliary information in estimating the population parameters by introducing ratio-type estimators. Later on, Das and Tripathi [6] suggested using the auxiliary information in the estimation of the variance of a finite population. Isaki [7] introduced a ratio estimator for the variance based on an auxiliary variable. Singh et al. [8] analyzed calibration estimators for the population variance. Kadilar and Cingi [9] used auxiliary variable to develop improved estimators of the variance of population. The new estimators achieved a significant improvement in efficiency over previous estimators. Subramani and Kumarapandiyan [10] suggested a variance estimator based on the median of the supplementary variable.

Gupta et al. [11] developed a generalization of the estimators of the variance under a linear scrambling model. The findings of the study suggested that the new estimator was more precise than the Isaki's [7] estimator under the linear response model. Recently, Saleem et al. [12] developed a new randomized response model and presented a novel estimator of the variance based on two auxiliary variables. The new estimator gained a significant boost in efficiency over the available variance estimators.

Zaman and Bulut [13] utilized the mathematical function of the auxiliary variable parameters to develop new variance estimators. In another study, Zaman and Bulut [14] suggested new estimators of variance by using the minimum covariance determinant. Another recent study of Zaman et al. [15] suggests a randomized technique for improving the efficiency of estimates using group discussion.

Using a new scrambling model, this paper presents an efficient variance estimator based on a non-sensitive ancillary variable. The suggested estimator is found to achieve a big improvement in efficiency over the Gupta et al. [11] estimator, and the Isaki [7] estimator under the proposed randomized response model.

This paper is outlined as follows:

Section 2 presents the mathematical notations and assumptions which have been used in the subsequent sections. Section 3 presents the mathematical equations of some of the available estimators from the literature. Section 4 presents the proposed variance estimator and its algebraic properties under the proposed randomized response model. Section 5 provides the performance evaluation measures for the proposed variance estimator. In Section 6, the performance of the suggested estimator is compared with the existing estimators and the results have been presented in tables. Section 7 presents a detailed discussion related to the findings of the analysis and the conclusion of the study.

2. Notations

Suppose the population of interest contains N units $U_1, U_2, \ldots, U_N$ and consider a simple random sample of n units obtained from a target population. Let the sensitive-type variable under consideration be denoted by Y, with X being the notation for an auxiliary variable having positive correlation with variable Y. Further, (x_i, y_i) denotes the value of (X, Y) corresponding to the ith unit of the population. The mathematical expressions for the different parameters and their estimators have been provided in Table 1.

3. Variance estimators in literature

A simple unbiased estimator of the variance may be expressed as:

$$t_0 = s_y^2. \tag{1}$$

The sampling variance of the estimator in equation (1) may be expressed as:

$$Var(t_0) = \theta S_y^4(\lambda_{40} - 1), \tag{2}$$

where $\theta = \frac{1}{n}$, and λ_{40} in equation (2) can be obtained from the general moment ratio λ_{rs} presented in Table 1.

Diana and Perri [16] proposed a linear model as:

$$Z = TY + S, \tag{3}$$

where Z is the response observed by the interviewer, and T and S denote scrambling/random variables, such that $E(T) = 1$ and $E(S) = 0$.

Under the scrambling model given in equation (3), the variance of Z can be derived as:

$$S_z^2 = S_{TY}^2 + S_S^2,$$

$$= S_T^2 S_y^2 + S_T^2 \mu_y^2 + S_y^2 + S_S^2,$$

or

$$S_y^2 = \frac{S_z^2 - S_S^2 - S_T^2 \overline{Z}^2}{S_T^2 + 1}. \tag{4}$$

If the variable of interest Y is sensitive, a randomized variant of the variance estimator t_0 can be obtained by replacing S_z^2 and $\overline{Z}^2$ by s_z^2 and $\overline{z}^2$, respectively. From equation (4), we get the following basic estimator of the variance:

$$t_0(R) = \frac{s_z^2 - S_S^2 - S_T^2 \overline{z}^2}{S_T^2 + 1}. \tag{5}$$

The estimator in equation (5) can be used in the development of ratio estimators of population variance.

Isaki [7] presented a ratio estimator for the population variance, which can be expressed as:

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

The bias and Mean Square Error (MSE) of the estimator given in equation (6), up to the first order of approximation, can be expressed as:

$$Bias(t_1) = \theta S_y^2(\lambda_{04} - \lambda_{22}), \tag{7}$$

and

$$MSE(t_1) = \theta S_y^4(\lambda_{40} + \lambda_{04} - 2\lambda_{22}). \tag{8}$$

The moment ratios λ_{40}, λ_{04}, and λ_{22} in equation (7) and equation (8) can be obtained from the general expression of λ_{rs} presented in Table 1.

Using the linear model, Gupta et al. [11] proposed the following generalized variance estimator:

$$t_2(R) = \left[\left(\left(\frac{s_z^2 - S_S^2 - S_T^2 \overline{z}^2}{S_T^2 + 1} \right) + \left(S_x^2 - s_x^2 \right) \right) \left(\frac{\alpha S_x^2 + \beta}{w\left(\alpha s_x^2 + \beta\right) + (1 - w)\left(\alpha S_x^2 + \beta\right)} \right)^g \right], \tag{9}$$

where g, w, α, and β are predetermined constants. For different values of constants, we can obtain special cases of the estimator given in equation (9).

The bias in the estimator $t_2(R)$ up to the first order of approximation may be derived as:

$$Bias(t_2(R)) = \frac{-\theta S_T^2 \overline{Z}^2}{S_T^2 + 1} C_z^2 - \frac{\alpha g w \theta S_x^2}{\alpha S_x^2 + \beta} \left[\frac{S_z^2(\lambda_{22} - 1) - 2S_T^2 \overline{Z}^2 \lambda_{12} C_z}{S_T^2 + 1} - S_x^2(\lambda_{04} - 1) \right], \tag{10}$$

In equation (10), the notation C_z^2 can be expressed as:

$$C_z^2 = C_y^2 S_T^2 + \frac{S_S^2}{\overline{Y}^2}.$$

The optimum MSE of $t_2(R)$ may be expressed as:

$$MSE(t_2(R))_{opt} = \frac{\theta}{\left(S_T^2 + 1\right)^2} \left[\left(S_z^4(\lambda_{40} - 1) + 4S_T^4 \overline{Z}^4 C_z^2 - 4S_z^2 S_T^2 \overline{Z}^2 \lambda_{30} C_z \right) \right.$$

$$-\frac{1}{(\lambda_{04}-1)}\left(S_z^2(\lambda_{22}-1)-2S_T^2\overline{Z}^2\lambda_{12}C_z\right)^2\Bigg]. \tag{11}$$

The expression for optimum MSE in equation (11) can be used for the purpose of efficiency comparison. Saleem et al. [12] proposed a generalized estimator of population variance as:

$$t_g=\left[k_1\left(\frac{s_z^2-S_S^2-S_T^2\overline{z}^2}{S_T^2+1}\right)+k_2\left(S_{x1}^2-s_{x1}^2\right)+k_3\left(S_{x2}^2-s_{x2}^2\right)\right]\exp\left(\frac{S_{x1}^2-s_{x1}^2}{S_{x1}^2+s_{x1}^2}\right)^{\alpha_1}\left(\frac{S_{x2}^2}{s_{x2}^2}\right)^{\alpha_2}, \tag{12}$$

where k_1, k_2, k_3, α_1, and α_2 are generalizing constants, (S_{x1}^2, S_{x2}^2) and (s_{x1}^2, s_{x2}^2) denote the population and sample variances of the auxiliary variables X_1 and X_2. Saleem et al. [12] also discussed many special cases of the estimator presented in equation (12). Two special cases of the Saleem et al. [12] generalized estimator are as follows:

$$t_3(R)=\frac{s_z^2-S_S^2-S_T^2\overline{z}^2}{S_T^2+1}\exp\left(\frac{S_x^2-s_x^2}{S_x^2+s_x^2}\right). \tag{13}$$

and

$$t_4(R)=\left[\frac{s_z^2-S_S^2-S_T^2\overline{z}^2}{S_T^2+1}+\left(S_x^2-s_x^2\right)\right]\exp\left(\frac{S_x^2-s_x^2}{S_x^2+s_x^2}\right). \tag{14}$$

Before deriving the mean squared error of each of the available estimators under the proposed model, we first present our proposed model in Section 4. Using our proposed model, we have derived the mean square error of various estimators, including those given in equation (13) and equation (14), in Section 5.

4. Proposed model and variance estimator

To estimate the population variance, the following scrambling model is proposed:

$$Z=\gamma(Y+S)+(1-\gamma)(Y+YS), \tag{15}$$

where γ is a constant such that $0\le\gamma\le1$, and S is a scrambling variable such that $E(S)=0$, and $Var(S)=S_S^2$. For our proposed model presented in equation (15), we assume that the sensitive variable Y is uncorrelated with the scrambling variable S.

Model's simplicity is one of the criteria for the practical usability of any randomized response model. In almost every sample survey using a randomized response technique, the respondent calculates his/her scrambled response and report it to the interviewer. This makes it necessary that the model employed for data collection should be simple enough so that the respondent can easily calculate and report his/her scrambled response. The proposed model uses a single scrambling variable and hence is simpler than many of the available randomized response models where two scrambling are used. Compared to a two-variable model, a one-variable model puts less burden on the respondents to calculate the scrambled response.

Using the proposed model, a simple estimator of the population variance may be obtained as:

$$S_z^2=\gamma^2S_{Y+S}^2+(1-\gamma)^2S_{Y+YS}^2. \tag{16}$$

Equation (16) can be further simplified as:

$$S_z^2=\gamma^2\left(S_y^2+S_S^2\right)+(1-\gamma)^2\left[S_y^2+E(YS)^2-\{E(YS)\}^2\right].$$

Using the independence of Y and S, further simplification yields:

$$S_z^2=\gamma^2\left(S_y^2+S_S^2\right)+(1-\gamma)^2\left[S_y^2+\left(S_y^2+\mu_y^2\right)S_S^2\right],$$

or

$$S_z^2=AS_y^2+\left[\gamma^2+(1-\gamma)^2\overline{Z}^2\right]S_S^2, \tag{17}$$

where

$$A=\gamma^2+(1-\gamma)^2+(1-\gamma)^2S_S^2. \tag{18}$$

Further simplification of equation (17) yields:

$$S_y^2=\frac{S_z^2-\left[\gamma^2+(1-\gamma)^2\overline{Z}^2\right]S_S^2}{A}, \tag{19}$$

where A is defined in equation (18).

In equation (19), replacing $\overline{Z}$ and S_z^2 by their unbiased estimators, $\overline{z}$ and s_z^2, respectively, we obtain the basic estimator of S_y^2 as:

$$t_0(R) = \frac{s_z^2 - \left[\gamma^2 + (1-\gamma)^2\overline{z}^2\right]S_S^2}{A}. \tag{20}$$

The estimator given in equation (20) can be used to develop new estimators of population variance. Motivated by the study of Azeem and Hanif [17], the following estimator of the population variance is proposed:

$$t_p = s_y^2 \frac{s_x^{*2}}{S_x^2}, \tag{21}$$

where

$$s_x^{*2} = \frac{(N-1)S_x^2 - (n-1)s_x^2}{N-n}, \tag{22}$$

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

and

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2.$$

Using the proposed model and using equation (22), the suggested estimator given in equation (21) may be expressed as:

$$t_p(R) = \frac{s_z^2 - \left[\gamma^2 + (1-\gamma)^2\overline{z}^2\right]S_S^2}{A}\left[\frac{(N-1)S_x^2 - (n-1)s_x^2}{(N-n)S_x^2}\right]. \tag{23}$$

In the subsequent theorems, we prove the algebraic properties of our proposed estimator presented in equation (23).

Theorem 1. *The bias of the proposed estimator may be expressed as*:

$$Bias\left(t_p(R)\right) \approx \frac{\theta}{A}\left[-(1-\gamma)^2\overline{Z}^2S_S^2C_Z^2 - DS_Z^2(\lambda_{22}-1) + 2D(1-\gamma)^2\overline{Z}^2S_S^2\lambda_{12}C_Z\right], \tag{24}$$

where

$$A = \gamma^2 + (1-\gamma)^2 + (1-\gamma)^2S_S^2, \text{and } D = \frac{n-1}{N-n}.$$

Proof: *In order to obtain the bias, let*

$$s_z^2 = S_z^2(1+d_z), s_x^2 = S_x^2(1+d_x), \text{and } \overline{z} = \overline{Z}(1+e_z),$$

where

$$d_z = \frac{s_z^2 - S_z^2}{S_z^2}, d_x = \frac{s_x^2 - S_x^2}{S_x^2}, \text{and } e_z = \frac{\overline{z} - \overline{Z}}{\overline{Z}},$$

so that

$$E(d_z) = E(d_x) = E(e_z) = 0, E\left(d_z^2\right) = \theta(\lambda_{40}-1), E\left(d_x^2\right) = \theta(\lambda_{04}-1), E\left(e_z^2\right) = \theta C_z^2,$$

$$E(d_zd_x) = \theta(\lambda_{22}-1), E(d_ze_z) = \theta\lambda_{30}C_z, \text{and } E(d_xe_z) = \theta\lambda_{12}C_z.$$

Using these notations, the proposed estimator may be expressed as:

$$t_p(R) = \frac{S_z^2(1+d_z) - \left\{\gamma^2 + (1-\gamma)^2\overline{Z}^2(1+e_z)^2\right\}S_S^2}{A}\left[\frac{(N-1)S_x^2 - (n-1)S_x^2(1+d_x)}{(N-n)S_x^2}\right],$$

or

$$t_p(R) = \frac{S_z^2 + S_z^2d_z - \left\{\gamma^2 + (1-\gamma)^2\overline{Z}^2\right\}S_S^2 - 2(1-\gamma)^2\overline{Z}^2S_S^2e_z - (1-\gamma)^2\overline{Z}^2S_S^2e_z^2}{A}$$

$$\times\left[\frac{(N-n)S_x^2 - (n-1)S_x^2d_x}{(N-n)S_x^2}\right],$$

or

$$t_p(R) = \frac{S_z^2 + S_z^2 d_z - \{\gamma^2 + (1-\gamma)^2\overline{Z}^2\}S_S^2 - 2(1-\gamma)^2\overline{Z}^2 S_S^2 e_z - (1-\gamma)^2\overline{Z}^2 S_S^2 e_z^2}{A}(1 - Dd_x),$$

where

$$D = \frac{n-1}{N-n}.$$

Further simplification yields:

$$t_p(R) - S_y^2 = \frac{S_z^2 d_z - 2(1-\gamma)^2\overline{Z}^2 S_S^2 e_z - (1-\gamma)^2\overline{Z}^2 S_S^2 e_z^2}{A} - \frac{Dd_x}{A}\left[S_z^2 + S_z^2 d_z - \{\gamma^2 + (1-\gamma)^2\overline{Z}^2\}S_S^2\right.$$

$$\left. -2(1-\gamma)^2\overline{Z}^2 S_S^2 e_z - (1-\gamma)^2\overline{Z}^2 S_S^2 e_z^2\right]. \tag{25}$$

Applying expectation and simplification yields:

$$E\left[t_p(R) - S_y^2\right] = \frac{1}{A}E\left[-(1-\gamma)^2\overline{Z}^2 S_S^2 e_z^2 - DS_z^2 d_z d_x + 2D(1-\gamma)^2\overline{Z}^2 S_S^2 d_x e_z\right].$$

On further simplification, we get the result given in equation (24) *as*:

$$Bias\left(t_p(R)\right) = \frac{\theta}{A}\left[-(1-\gamma)^2\overline{Z}^2 S_S^2 C_z^2 - DS_z^2(\lambda_{22} - 1) + 2D(1-\gamma)^2\overline{Z}^2 S_S^2 \lambda_{12} C_z\right].$$

This completes the proof.

Theorem 2. *The MSE of the suggested estimator can be obtained as*:

$$MSE\left(t_p(R)\right) = \frac{\theta}{A^2}\left[S_z^4(\lambda_{40} - 1) + 4(1-\gamma)^4\overline{Z}^4 S_S^4 C_z^2 + B^2D^2(\lambda_{04} - 1) - 4(1-\gamma)^2\overline{Z}^2 S_z^2 S_S^2 \lambda_{30} C_z\right.$$

$$\left. -2BDS_z^2(\lambda_{22} - 1) + 4(1-\gamma)^2 BD\overline{Z}^2 S_S^2 \lambda_{12} C_z\right],$$

where

$$B = S_z^2 - \{\gamma^2 + (1-\gamma)^2\overline{Z}^2\}S_S^2.$$

Proof:
Ignoring higher order terms, equation (25) *simplifies to*:

$$t_p(R) - S_y^2 = \frac{S_z^2 d_z - 2(1-\gamma)^2\overline{Z}^2 S_S^2 e_z}{A} - \frac{Dd_x}{A}\left[S_z^2 - \{\gamma^2 + (1-\gamma)^2\overline{Z}^2\}S_S^2\right],$$

or

$$t_p(R) - S_y^2 = \frac{1}{A}\left[S_z^2 d_z - 2(1-\gamma)^2\overline{Z}^2 S_S^2 e_z - DS_z^2 d_x + D\{\gamma^2 + (1-\gamma)^2\overline{Z}^2\}S_S^2 d_x\right]. \tag{26}$$

Squaring both sides of equation (26) *and applying expectation, we get*:

$$E\left[t_p(R) - S_y^2\right]^2 = \frac{1}{A^2}E\left[S_z^2 d_z - 2(1-\gamma)^2\overline{Z}^2 S_S^2 e_z - BDd_x\right]^2,$$

where

$$B = S_z^2 - \{\gamma^2 + (1-\gamma)^2\overline{Z}^2\}S_S^2.$$

Further simplification yields the required result.

$$MSE\left(t_p(R)\right) = \frac{\theta}{A^2}\left[S_z^4(\lambda_{40} - 1) + 4(1-\gamma)^4\overline{Z}^4 S_S^4 C_z^2 + B^2D^2(\lambda_{04} - 1) - 4(1-\gamma)^2\overline{Z}^2 S_z^2 S_S^2 \lambda_{30} C_z\right.$$

$$\left. -2BDS_z^2(\lambda_{22} - 1) + 4(1-\gamma)^2 BD\overline{Z}^2 S_S^2 \lambda_{12} C_z\right]. \tag{27}$$

Differentiating equation (27) *with respect to* γ *and then equating to zero yields the optimum value as*:

$$\gamma_{opt} = 1 - \sqrt{\frac{S_S^2\left(S_z^2\lambda_{30} - BD\lambda_{12}\right)}{2\overline{Z}^2 S_z^4 C_z}}. \tag{28}$$

The optimum value of γ *from* equation (28) *may be used in* equation (27) *to get the optimum mean squared error of the suggested estimator.*

5. Performance evaluation

Yan et al. [18] proposed a novel metric to quantify the degree of the respondents' privacy. In the case of our proposed model, the respondents' degree of privacy can be calculated as:

$$\Delta = E(Z-Y)^2 = E[\gamma S + (1-\gamma)YS]^2,$$

$$= \gamma^2 E(S^2) + (1-\gamma)^2 E(Y^2S^2) + 2\gamma(1-\gamma)E(YS^2),$$

$$= E(S^2)\left[\gamma^2 + (1-\gamma)^2 E(Y^2) + 2\gamma(1-\gamma)E(Y)\right],$$

or

$$\Delta = S_S^2\left[\gamma^2 + (1-\gamma)^2\left(S_y^2 + \mu_y^2\right) + 2\gamma(1-\gamma)\mu_y\right]. \tag{29}$$

The measure Δ given in equation (29) uses the respondents' privacy protection level but ignores another important aspect of model quality – its efficiency. A better approach would be to quantify privacy and efficiency into a single metric. For this purpose, a unified metric was proposed by Gupta et al. [19] as:

$$\delta = \frac{MSE}{\Delta}. \tag{30}$$

The model-evaluation measure given in equation (30) is useful for comparison of models.

Using the proposed model, the mathematical expression for the bias and the mean squared error of the basic estimator $t_0(R)$ may be obtained as:

$$Bias(t_0(R)) = \frac{-\theta\overline{Z}^2(1-\gamma)^2 S_S^2 C_z^2}{A}, \tag{31}$$

and

$$MSE(t_0(R)) = \frac{\theta}{A^2}\left[S_z^4(\lambda_{40}-1) + 4(1-\gamma)^4\overline{Z}^4 S_S^4 C_z^2 - 4(1-\gamma)^2\overline{Z}^2 S_z^2 S_S^2 \lambda_{30} C_z\right]. \tag{32}$$

The symbol A in equation (31) and equation (32) has been defined in equation (18). The bias and mean squared error of the Isaki's [7] estimator $t_1(R)$ under the proposed model may be derived as:

$$Bias(t_1(R)) = \frac{-\theta}{A}\left[(1-\gamma)^2\overline{Z}^2 S_S^2 C_z^2 + S_z^2(\lambda_{22}-1) - 2(1-\gamma)^2\overline{Z}^2 S_S^2 \lambda_{12} C_z - B(\lambda_{04}-1)\right], \tag{33}$$

and

$$MSE(t_1(R)) = \frac{\theta}{A^2}\left[S_z^4(\lambda_{40}-1) + 4(1-\gamma)^4\overline{Z}^4 S_S^4 C_z^2 + B^2(\lambda_{04}-1) - 4(1-\gamma)^2\overline{Z}^2 S_S^2 S_z^2 \lambda_{30} C_z\right.$$

$$\left. -2BS_z^2(\lambda_{22}-1) + 4(1-\gamma)^2 B\overline{Z}^2 S_S^2 \lambda_{12} C_z\right]. \tag{34}$$

The bias and MSE of the Gupta et al. [11] estimator $t_2(R)$ under the proposed model may be derived as:

$$Bias(t_2(R)) = \frac{\theta}{A}\left[-(1-\gamma)^2\overline{Z}^2 S_S^2 C_z^2 - gCS_z^2(\lambda_{22}-1) + 2(1-\gamma)^2 gC\overline{Z}^2 S_S^2 \lambda_{12} C_z\right.$$

$$\left. + \frac{g(g+1)}{2} C^2 B(\lambda_{04}-1)\right], \tag{35}$$

and

$$MSE(t_2(R)) = \frac{\theta}{A^2}\left[S_z^4(\lambda_{40}-1) + 4(1-\gamma)^4\overline{Z}^4 S_S^4 C_z^2 + \left(AS_x^2 + gBC\right)^2(\lambda_{04}-1) - 4(1-\gamma)^2\overline{Z}^2 S_z^2 S_S^2 \lambda_{30} C_z\right.$$

$$\left. + 4(1-\gamma)^2\overline{Z}^2\left(AS_x^2 + gBC\right)S_S^2\lambda_{12}C_z - 2\left(AS_x^2 + gBC\right)S_z^2(\lambda_{22}-1)\right], \tag{36}$$

where

$$C = \frac{w\alpha S_x^2}{\alpha S_x^2 + \beta}.$$

The optimum value of g may be obtained by differentiating equation (36) with respect to g and solution of the equation yields:

Table 2
MSEs of various estimators for $S_x^2 = 10$, $S_z^2 = 8$, $\mu_y = 3$, $S_y^2 = 2$, $N = 5000$, $\alpha = 2$, $\beta = 3$.

γ	w	n	$t_1(R)$	$t_2(R)$	$t_3(R)$	$t_4(R)$	$t_p(R)$
0.8	0.3	50	24.9205	45.1686	18.6367	36.7023	17.0639
		100	12.4603	22.5843	9.3183	18.3512	8.5251
		200	6.2301	11.2921	4.6592	9.1756	4.2571
		500	2.4921	4.5169	1.8637	3.6702	1.7017
		1000	1.2460	2.2584	0.9318	1.8351	0.8633
	0.8	50	24.9539	103.9532	18.6438	36.7635	17.0453
		100	12.4770	51.9766	9.3219	18.3817	8.5161
		200	6.2385	25.9883	4.6609	9.1909	4.2529
		500	2.4954	10.3953	1.8644	3.6763	1.7003
		1000	1.2477	5.1977	0.9322	1.8382	0.8630
0.4	0.3	50	14.7725	34.6525	11.9102	30.2480	10.4305
		100	7.3862	17.3262	5.9551	15.1240	5.2240
		200	3.6931	8.6631	2.9775	7.5620	2.6216
		500	1.4772	3.4652	1.1910	3.0248	1.0628
		1000	0.7386	1.7326	0.5955	1.5124	0.5495
	0.8	50	14.9552	61.6075	11.9615	30.8018	10.3531
		100	7.4776	30.8038	5.9808	15.4009	5.1867
		200	3.7388	15.4019	2.9904	7.7004	2.6043
		500	1.4955	6.1608	1.1962	3.0802	1.0577
		1000	0.7478	3.0804	0.5981	1.5401	0.5488

Table 3
δ values of various estimators for $S_x^2 = 10$, $S_z^2 = 8$, $\mu_y = 3$, $S_y^2 = 2$, $N = 5000$, $\alpha = 2$, $\beta = 3$.

γ	w	n	$t_1(R)$	$t_2(R)$	$t_3(R)$	$t_4(R)$	$t_p(R)$
0.8	0.3	50	12.2159	22.1414	9.1356	17.9913	8.3646
		100	6.1080	11.0707	4.5678	8.9957	4.1790
		200	3.0540	5.5354	2.2839	4.4978	2.0868
		500	1.2216	2.2141	0.9136	1.7991	0.8341
		1000	0.6108	1.1071	0.4568	0.8996	0.4232
	0.8	50	12.2323	50.9575	9.1391	18.0213	8.3555
		100	6.1162	25.4787	4.5696	9.0107	4.1745
		200	3.0581	12.7394	2.2848	4.5053	2.0847
		500	1.2232	5.0957	0.9139	1.8021	0.8335
		1000	0.6116	2.5479	0.4570	0.9011	0.4230
0.4	0.3	50	2.6569	6.2325	2.1421	5.4403	1.8760
		100	1.3285	3.1162	1.0711	2.7201	0.9396
		200	0.6642	1.5581	0.5355	1.3601	0.4715
		500	0.2657	0.6232	0.2142	0.5440	0.1912
		1000	0.1328	0.3116	0.1071	0.2720	0.0988
	0.8	50	2.6898	11.0805	2.1514	5.5399	1.8621
		100	1.3449	5.5402	1.0757	2.7699	0.9329
		200	0.6724	2.7701	0.5378	1.3850	0.4684
		500	0.2690	1.1080	0.2151	0.5540	0.1902
		1000	0.1345	0.5540	0.1076	0.2770	0.0987

$$\widehat{g}_{opt} = \frac{S_z^2(\lambda_{22}-1) - 2(1-\gamma)^2\overline{Z}^2 S_S^2\lambda_{12}C_z - AS_x^2(\lambda_{04}-1)}{BC(\lambda_{04}-1)}. \tag{37}$$

This optimum value of g from equation (37) may be used in equation (36) to get the optimum variance of the Gupta et al. [11] estimator $t_2(R)$.

The MSE of the Saleem et al. [12] estimator $t_3(R)$ under the proposed model may be derived as:

$$MSE(t_3(R)) = \frac{\theta}{4A^2}\left[4S_z^4(\lambda_{40}-1) + 16(1-\gamma)^4\overline{Z}^4S_S^4C_z^2 + B^2(\lambda_{04}-1) - 16(1-\gamma)^2\overline{Z}^2S_z^2S_S^2\lambda_{30}C_z\right.$$

$$\left. -4BS_z^2(\lambda_{22}-1) + 8(1-\gamma)^2B\overline{Z}^2S_S^2\lambda_{12}C_z\right]. \tag{38}$$

The MSE of the Saleem et al. [12] estimator $t_4(R)$ under the proposed model may be derived as:

$$MSE(t_4(R)) = \frac{\theta}{A^2}\left[S_z^4(\lambda_{40}-1) + 4(1-\gamma)^4\overline{Z}^4S_S^4C_z^2 + E^2(\lambda_{04}-1) - 4(1-\gamma)^2\overline{Z}^2S_z^2S_S^2\lambda_{30}C_z\right.$$

$$\left. -2ES_z^2(\lambda_{22}-1) + 4(1-\gamma)^2E\overline{Z}^2S_S^2\lambda_{12}C_z\right], \tag{39}$$

Table 4
Root Mean Square Error (RMSE) of various estimators for $S_x^2 = 10$, $S_z^2 = 8$, $\mu_y = 3$, $S_y^2 = 2$, $N = 5000$, $\alpha = 2$, $\beta = 3$.

γ	w	n	$t_1(R)$	$t_2(R)$	$t_3(R)$	$t_4(R)$	$t_p(R)$
0.8	0.3	50	4.9920	6.7208	4.3170	6.0582	4.1308
		100	3.5299	4.7523	3.0526	4.2838	2.9198
		200	2.4960	3.3604	2.1585	3.0291	2.0633
		500	1.5786	2.1253	1.3652	1.9158	1.3045
		1000	1.1163	1.5028	0.9653	1.3547	0.9291
	0.8	50	4.9954	10.1957	4.3178	6.0633	4.1286
		100	3.5323	7.2095	3.0532	4.2874	2.9182
		200	2.4977	5.0979	2.1589	3.0316	2.0622
		500	1.5797	3.2242	1.3654	1.9174	1.3040
		1000	1.1170	2.2798	0.9655	1.3558	0.9290
0.4	0.3	50	3.8435	5.8866	3.4511	5.4998	3.2296
		100	2.7178	4.1625	2.4403	3.8890	2.2856
		200	1.9217	2.9433	1.7256	2.7499	1.6191
		500	1.2154	1.8615	1.0913	1.7392	1.0309
		1000	0.8594	1.3163	0.7717	1.2298	0.7413
	0.8	50	3.8672	7.8490	3.4585	5.5499	3.2176
		100	2.7345	5.5501	2.4456	3.9244	2.2774
		200	1.9336	3.9245	1.7293	2.7750	1.6138
		500	1.2229	2.4821	1.0937	1.7550	1.0285
		1000	0.8647	1.7551	0.7734	1.2410	0.7408

where

$$E = AS_x^2 + \frac{B}{2}.$$

We have used the results presented in equation 33–35 and equation 38 and 39 for comparison of the models in Table 2 and Table 3.

6. Comparison of estimators

Our suggested estimator will be more efficient than the basic estimator $t_0(R)$ if:

$$MSE\big(t_p(R)\big) < MSE(t_0(R)),$$

or

$$BD(\lambda_{04} - 1) < 2S_z^2(\lambda_{22} - 1) - 4(1-\gamma)^2\overline{Z}^2S_S^2\lambda_{12}C_z.$$

Our suggested estimator will be more efficient than the Isaki's [7] estimator $t_1(R)$ if:

$$MSE\big(t_p(R)\big) < MSE(t_1(R)),$$

or

$$n - 1 < N - n.$$

This condition is strong and always holds if the population size is more than twice the sample size.
The proposed estimator $t_p(R)$ will be more efficient than the Gupta et al. [11] estimator $t_2(R)$ if:

$$MSE\big(t_p(R)\big) < MSE(t_2(R)),$$

or

$$AS_x^2 + B(gC - 1) > 0.$$

The proposed estimator $t_p(R)$ will be more efficient than the Saleem et al. [12] estimator $t_3(R)$ if:

$$MSE\big(t_p(R)\big) < MSE(t_3(R)).$$

On simplification, the above condition reduces to:

$$B < \frac{2(D-2)}{(D^2-1)(\lambda_{04}-1)}\left[S_z^2(\lambda_{22}-1) - 2(1-\gamma)^2\overline{Z}^2S_S^2\lambda_{12}C_z\right].$$

The proposed estimator $t_p(R)$ will be more efficient than the Saleem et al. [12] estimator $t_4(R)$ if:

Table 5
Simulated Mean Absolute Error (MAE) of various estimators for $S_x^2 = 10$, $N = 5000$, $\alpha = 2$, $\beta = 3$, $g = 5$.

γ	w	n	$t_1(R)$	$t_2(R)$	$t_3(R)$	$t_4(R)$	$t_p(R)$
0.8	0.3	50	10.4141	10.2237	10.0333	10.0536	9.9678
		100	10.2323	10.1426	10.0490	10.0620	10.0175
		200	10.1975	10.1536	10.1284	10.1309	10.1215
		500	10.1625	10.1455	10.1287	10.1315	10.1223
		1000	10.1909	10.1836	10.1798	10.1801	10.1768
	0.8	50	10.4141	11.9726	10.0333	10.0536	9.9678
		100	10.2323	10.9741	10.0490	10.0620	10.0175
		200	10.1975	10.5047	10.1284	10.1309	10.1215
		500	10.1625	10.3033	10.1287	10.1315	10.1223
		1000	10.1909	10.2415	10.1798	10.1801	10.1768
0.4	0.3	50	6.1595	6.0622	5.9214	5.9341	5.8770
		100	4.7969	4.7726	4.6920	4.7071	4.6769
		200	4.1433	4.1368	4.1057	4.1104	4.1027
		500	3.7233	3.7169	3.7119	3.7119	3.7113
		1000	3.7072	3.7047	3.7016	3.7021	3.7000
	0.8	50	6.1595	7.1953	5.9214	5.9341	5.8770
		100	4.7969	5.3234	4.6920	4.7071	4.6770
		200	4.1433	4.3700	4.1057	4.1104	4.1027
		500	10.4141	10.2237	10.0333	10.0536	9.9678
		1000	10.2323	10.1426	10.0490	10.0620	10.0174

$$MSE\big(t_p(R)\big) < MSE(t_4(R)).$$

On simplification, the above condition reduces to:

$$BD + E < \frac{2}{\lambda_{04} - 1}\left[S_z^2(\lambda_{22} - 1) - 2(1-\gamma)^2\overline{Z}^2 S_S^2 \lambda_{12} C_z\right].$$

The mean square errors of the estimators $t_0(R)$, $t_1(R)$, $t_2(R)$, $t_3(R)$, $t_4(R)$, and the suggested estimator $t_p(R)$ are displayed in Table 2 for various values of w, γ, and the sample size n. Table 3 presents the δ values for various choices of S_T^2, S_S^2, and the sample size n. Table 4 presents the values of root mean square error (RMSE) of various estimators. Examining Table 2 to Table 4, the improvement in terms of efficiency and δ values may clearly be observed. Moreover, Table 5 shows the simulated values of Mean Absolute Error (MAE) of the proposed and other variance estimators, based on 1000 iterations using different sample sizes from an artificial population generated through R code. The improvement in terms of mean absolute deviation can also be observed from Table 5.

7. Discussion and conclusion

This study introduced a new randomized response model for precise estimation of the variance of a finite population. Additionally, a new estimator of the variance has been developed which outperforms the existing variance estimators. The mathematical properties of the suggested variance estimator under the proposed model have been derived. Table 2 shows the mean square error of the Isaki [7] estimator, the Gupta et al. [11] estimator, and the suggested estimator for different sample sizes and for various choices of the constants. The corresponding δ values have been presented in Table 3 for different sample sizes under the proposed model.

Table 2 clearly shows that, based on the proposed model, the proposed estimator is the most efficient estimator. It may also be examined in the table that an increase in the sample size n results in a decline in the mean square error of each estimator. It is also clear that the Isaki's [7] estimator performs better than the Gupta et al. [11] estimator under the proposed model. It may also be observed that as the value of γ changes from 0.8 to 0.4, the mean square error of each estimator decreases.

Glancing at the combined measure of estimator quality, δ, presented in Table 3, one may observe that the proposed estimator produces the best δ values of all three estimators. This makes the suggested variance estimator the most suitable estimator for use with sensitive surveys. Table 3 also shows that the Isaki's [7] estimator produces smaller δ values than the Gupta et al. [11] estimator under the proposed model. Based on the findings of this study, it is recommended for survey researchers to use the proposed variance estimator in situations where the variable of interest is of sensitive nature. Table 5 shows that the proposed variance estimator achieves the least mean absolute error of all five estimators.

The proposed variance estimator is designed for use with simple random sampling where the variable of interest is sensitive in nature. It is recommended for future researchers to extend the proposed estimator and/or the proposed model to other sampling schemes, including stratified sampling and systematic sampling. The proposed estimator can also be used in unequal probability sampling, and it is therefore suggested that future researchers analyze its properties under unequal probability sampling.

It may also be interesting if future researchers analyze the properties of the new suggested estimator in the case of measurement error and non-response error. Researchers may also work on modifying the proposed model for even more improvement in efficiency.

Data availability

All relevant data is available within the manuscript.

Funding for the study

The authors received no funding for this study.

CRediT authorship contribution statement

Muhammad Azeem: Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Najma Salahuddin:** Writing – review & editing, Visualization, Software, Investigation, Data curation. **Sundus Hussain:** Writing – review & editing, Software, Methodology, Investigation. **Musarrat Ijaz:** Validation, Project administration, Data curation. **Abdul Salam:** Validation, Software, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S.L. Warner, Randomized response: a survey technique for eliminating evasive answer bias, J. Am. Stat. Assoc. 60 (309) (1965) 63–69, https://doi.org/10.1080/01621459.1965.10480775.
[2] S.L. Warner, The linear randomized response model, J. Am. Stat. Assoc. 66 (336) (1971) 884–888, https://doi.org/10.1080/01621459.1971.10482364.
[3] S. Gupta, B. Gupta, S. Singh, Estimation of sensitivity level of personal interview survey questions, J. Stat. Plann. Inference 100 (2) (2002) 239–247, https://doi.org/10.1016/S0378-3758(01)00137-9.
[4] M. Azeem, Using the exponential function of scrambling variable in quantitative randomized response models, Math. Methods Appl. Sci. 46 (13) (2023) 13882–13893, https://doi.org/10.1002/mma.9295.
[5] W.G. Cochran, The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce, J. Agric. Sci. 30 (1940) 262–275, https://doi.org/10.1017/S0021859600048012.
[6] A.K. Das, T.P. Tripathi, Use of auxiliary information in estimating the finite population variance, Sankhya 40 (1978) 139–148.
[7] C.T. Isaki, Variance estimation using auxiliary information, J. Am. Stat. Assoc. 78 (381) (1983) 117–123, https://doi.org/10.2307/2287117.
[8] S. Singh, S. Horn, S. Chowdhury, F. Yu, Calibration of the estimators of variance, Aust. N. Z. J. Stat. 41 (1999) 199–212, https://doi.org/10.1111/1467-842X.00074.
[9] C. Kadilar, H. Cingi, Improvement in variance estimation using auxiliary information, Hacettepe Journal of Mathematics and Statistics 35 (1) (2006) 111–115.
[10] J. Subramani, G. Kumarapandiyan, Variance estimation using median of the auxiliary variable, Int. J. Probab. Stat. 1 (3) (2012) 36–40, https://doi.org/10.5923/j.ijps.20120103.02.
[11] S. Gupta, B. Aloraini, M.N. Qureshi, S. Khalil, Variance estimation using randomized response technique, REVSTAT – Statistical Journal 18 (2) (2020) 165–176.
[12] I. Saleem, A. Sanaullah, L.A. Al-Essa, S. Bashir, S. Efficient estimation of population variance of a sensitive variable using a new scrambling response model, Sci. Rep. 13 (2023) 1–11, https://doi.org/10.1038/s41598-023-45427-2.
[13] T. Zaman, H. Bulut, A New Class of Robust Ratio Estimators for Finite Population Variance, Scientia Iranica, 2023, https://doi.org/10.24200/sci.2022.57175.5100.
[14] T. Zaman, H. Bulut, An efficient family of robust-type estimators for the population variance in simple and stratified random sampling, Commun. Stat. Theor. Methods 52 (8) (2023) 2610–2624, https://doi.org/10.1080/03610926.2021.1955388.
[15] Q. Zaman, M. Ijaz, T. Zaman, A randomization tool for obtaining efficient estimators through focus group discussion in sensitive surveys, Commun. Stat. Theor. Methods 52 (10) (2023) 3414–3428, https://doi.org/10.1080/03610926.2021.1973502.
[16] G. Diana, P.F. Perri, A class of estimators of quantitative sensitive data, Stat. Pap. 52 (3) (2011) 633–650, https://doi.org/10.1007/s00362-009-0273-1.
[17] M. Azeem, M. Hanif, Joint influence of measurement error and non-response on estimation of population mean, Commun. Stat. Theor. Methods 46 (4) (2017) 1679–1693, https://doi.org/10.1080/03610926.2015.1026992.
[18] Z. Yan, J. Wang, J. Lai, An efficiency and protection degree-based comparison among the quantitative randomized response strategies, Commun. Stat. Theor. Methods 38 (3) (2008) 400–408, https://doi.org/10.1080/03610920802220785.
[19] S. Gupta, S. Mehta, J. Shabbir, S. Khalil, A unified measure of respondent privacy and model efficiency in quantitative rrt models, Journal of Statistical Theory and Practice 12 (3) (2018) 506–511, https://doi.org/10.1080/15598608.2017.1415175.