On Generalized Additive Scrambled Response Modeling in Sensitive Surveys

Waqas ARSHAD1,* and Zawar HUSSAIN1

1Department of Statistics, Quaid-i-Azam University 45320, Islamabad, Pakistan

Abstract

In this article, we use additive scrambling to estimate the mean of a sensitive variable. In the proposed scrambling model, taking G (>1 ) as a positive integer chosen by the interviewer, each respondent is asked to randomly draw G values from a given distribution of scrambling variable and add average of these randomly drawn values to his/her true response on the sensitive variable. Using repetition of the scrambling experiment, we propose a relatively more efficient estimator of sensitive mean without incurring any additional sampling cost. We present a generalization of additive scrambled response models and show that most of additive scrambling models are special cases of suggested generalization. Through algebraic and numerical comparisons, superiority of the proposed methodology is established.

1. INTRODUCTION

Asking sensitive questions directly is always a tricky issue because respondents may feel insecure and become reluctant to divulge the true response. Sensitive questions are related to habitual tax evasion, abortion, drunken driving, gambling, drug abuse, number of bottles of vine consumed in a month, number of violations of traffic rules in a year, number of cheating in exam during the whole educational career, etc. Respondents mostly give evasive answers or refuse to give response on sensitive questions. [1] introduced a technique called Randomized Response Technique (RRT), consisting of two complementary questions; one of them is randomly chosen by the respondents and answered truthfully. The respondent only reports a ‘yes’ or ‘no’ and does not unveil the question randomly selected by him/her. Through this RRT, honest and reliable data are gathered and unbiased estimate of the proportion of individuals possessing sensitive attribute is obtained. This method proved very effective in medical, socioeconomic, biological and many other fields of life to collect trustworthy data.

After the pioneering work of [1], different contributions towards randomized response sampling have been put forward in the literature. [2] and [3,4] introduced an unrelated question technique consisting of two questions; a question about sensitive variable and a question about unrelated (non sensitive) variable. By using this technique, privacy of the respondent is more protected. [5-15] modified the work of [1] to improve the efficiency of the estimators and provided increased privacy protection. Some more work can be seen in the following papers of [16-21] etc.

There are three types of scrambled response models we found in the literature, namely, additive scrambled response models (ASRM), multiplicative scrambled response models (MSRM) and mixed scrambled response models. For each model to be discussed later, the reported response is denoted by Z. The corresponding estimators and their variances differ in the reported response Z, its sample and population means \( \bar{Z} \) and \( \mu_Z \) and its population variance \( \sigma_Z^2 \). The traditional ASRM was introduced by [22]. In this model, a scrambling variable (with known distribution) is added to the true sensitive response. This model
provides good privacy protection to the respondent and may be briefly explained as follows. Let $X$ be a study variable with mean $\mu_X$ and variance $\sigma_X^2$. Let $Y$ be a scrambling variable with mean $\mu_Y = 0$ and variance $\sigma_Y^2$. The respondent chosen by simple random sampling gives the additive scrambled response as $Z = X + Y$.

(1)

The unbiased estimator, $\hat{\mu}_{XH}$, of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XH} = \bar{Z},$$

(2)

and

$$\text{Var}(\hat{\mu}_{XH}) = \frac{\sigma_X^2 + \sigma_Y^2}{n}.$$  

(3)

The above model may also be termed as a zero stage model.

[23] proposed a two stage optional randomized response model by using a partial scrambling approach. Two subsamples are required in this approach. In each subsample, a proportion $P$ of respondents give the true response and remaining proportion $(1-P)$ of the respondents is directed to go to the second stage.

Every respondent, using second stage, has the option to report on the sensitive question if he/she feels the question insensitive or to report an additive scrambled response, otherwise. Let $W$ be the population proportion of individuals who feels the study question as sensitive. Let $X_i$ be the study variable with mean $\mu_X$ and variance $\sigma_X^2$. Let $Y_i$ be the scrambling variable with mean $\theta_i$, $i = 1, 2$ and variance $\sigma_Y^2$. The total sample size $n$ is divided into two subsamples of sizes $n_1$ and $n_2$. The distribution of the reported response in the $i^{th}$ subsample is given by

$$Z_i = \begin{cases} X & \text{with probability } P + (1-P)(1-W) \\ X + Y_i & \text{with probability } W(1-P) \end{cases}.$$  

(4)

An unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XG} = \frac{\theta_1 \bar{Z}_1 - \theta_1 \bar{Z}_2}{\theta_2 - \theta_1},$$

(5)

and

$$\text{Var}(\hat{\mu}_{XG}) = \frac{1}{(\theta_2 - \theta_1)^2} \left\{ \theta_2 \frac{\sigma_X^2}{n_1} + \theta_1 \frac{\sigma_Y^2}{n_2} \right\}.$$  

(6)

where $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 [(1-P)W] + \theta_1^2 [(1-P)W] [1 - (1-P)W]$.

[24] proposed a different two stage procedure. At 1st stage, a proportion $T$ of the respondents gives the scrambled response and the remaining proportion $(1-T)$ of the respondents has two options; either to give true response with probability $(1-W)$ or to give additive scrambled response with probability $W$, where $W$ is the sensitivity level. Let $X$ be the study variable with mean $\mu_X$ and variance $\sigma_X^2$. Let $Y_i$ be the scrambling variable with mean $\mu_Y$, $i = 1, 2$ and variance $\sigma_Y^2$. The total sample size $n$ is divided into two subsamples of sizes $n_1$ and $n_2$. The observed response $Z_i$ from the $i^{th}$ subgroup may be written as

$$Z_i = \begin{cases} X & \text{with probability } (1-T)(1-W) \\ X + Y_i & \text{with probability } T + W(1-T) \end{cases}.$$  

(7)

An unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XM} = \frac{\theta_2 \bar{Z}_1 - \theta_1 \bar{Z}_2}{\theta_2 - \theta_1},$$

(8)

and

$$\text{Var}(\hat{\mu}_{XM}) = \frac{1}{(\theta_2 - \theta_1)^2} \left\{ \theta_2 \frac{\sigma_X^2}{n_1} + \theta_1 \frac{\sigma_Y^2}{n_2} \right\}.$$  

(9)

where $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 [T + (1-T)W] + \theta_1^2 [T + (1-T)W] [1 - T + (1-T)W]$.

[21] proposed a scrambling model based on two random subsamples from the population. The respondents are requested to give the true response or to give a scrambled response. In this model, two scrambling variables are used. Let $Y_i$ be the first scrambling variable with mean $\mu_Y = 1$ and variance $\sigma_Y^2$. Let $X$ be the study variable with mean $\mu_X$ and variance $\sigma_X^2$. Let $S_i$ be the second scrambling variable with mean $\mu_S = \theta_i$ and variance $\sigma_S^2 = \gamma^2$. The total sample size $n$ is divided in two subsamples of sizes $n_1$ and $n_2$. The optional randomized response model for the $i^{th}$ subsample is given by

$$Z_i = (1-T)X + T(S_iX + Y_i).$$  

(10)
where \( T \sim \text{Bernoulli}(W) \) and \( W \) is sensitivity level. The unbiased estimator of population mean and its variance are given, respectively, by
\[
\hat{\mu}_{XHU} = \frac{\theta_1 z_2 - \theta_2 z_1}{\theta_1 - \theta_2} \tag{11}
\]
and
\[
\text{Var}(\hat{\mu}_{XHU}) = \frac{1}{(\theta_1 - \theta_2)^2} \left\{ \theta_1^2 \sigma_{z_2}^2 + \theta_2^2 \sigma_{z_1}^2 \right\} \tag{12}
\]
where \( \sigma_{z_i}^2 = \sigma_X^2 + W (\sigma_Y^2 + \mu_Y^2) + W Y_i^2 + W (1 - W) \theta_i^2 \).

[25] modified the [21] model and proposed a two stage model. At the 1\(^{st}\) stage the respondents have to give the true response with probability \( P \) or to move to second stage with probability \( 1 - P \). The 2\(^{nd}\) stage is the same as that of [21]. The response of the \( r^{th} \) respondent is given by
\[
Z_i = \begin{cases} 
X & \text{with probability } P (1 - P) (1 - W) \\
S_i X + Y_i & \text{with probability } W (1 - P) 
\end{cases} \tag{13}
\]
An unbiased estimator of population mean and variance are given, respectively, by
\[
\hat{\mu}_{XG2} = \frac{\theta_2 z_2 - \theta_1 z_1}{\theta_2 - \theta_1} \tag{14}
\]
and
\[
\text{Var}(\hat{\mu}_{XG2}) = \frac{1}{(\theta_2 - \theta_1)^2} \left\{ \theta_2^2 \sigma_{z_2}^2 + \theta_1^2 \sigma_{z_1}^2 \right\} \tag{15}
\]
where \( \sigma_{z_i}^2 = \sigma_X^2 + W (1 - P) \sigma_Y^2 (\sigma_X^2 + \mu_X^2) + W (1 - P) (Y_i^2 + \theta_i^2) - W^2 (1 - P^2) \theta_i^2 \).

MSRM was proposed by Eichhorn and Hayre (1983). In this model, the scrambling variable, having a known distribution, is multiplied by the true response. Their model increases the privacy protection. [7] gave a one stage MSRM where respondents have the option to report a true response or a scrambled response. If a respondent feels question as insensitive, he/she reports the true value of sensitive variable, otherwise, he/she reports a scrambled response.

Ryu et al. (2006) introduced a two stage MSRM. At 1\(^{st}\) stage, the respondents are given two statements; either they can choose the direct question with probability \( P \) or move to 2\(^{nd}\) stage with probability \( 1 - P \). At the 2\(^{nd}\) stage the respondents give answer to the direct question with known probability \( T \) or a scrambled response with probability \( 1 - T \). Let \( X \) be the study variable with mean \( \mu_X \) and variance \( \sigma_X^2 \). Let \( Y \) be the scrambling variable with mean \( \mu_Y \) and variance \( \sigma_Y^2 \). An unbiased estimator of population mean and its variance are given, respectively, by
\[
\hat{\mu}_X = \bar{Z} \tag{16}
\]
and
\[
\text{Var}(\hat{\mu}_X) = \frac{1}{n} \left\{ \sigma_Y^2 + (1 - P) T \sigma_Y^2 (\sigma_X^2 + \mu_X^2) \right\} \tag{17}
\]
[15] modified the Ryu et al. (2006) model by modifying the 2\(^{nd}\) stage. At 2\(^{nd}\) stage, respondents give the true response with probability \( W \) or give scrambled response with probability \( W \), where \( W \) is termed as sensitivity level of the sensitive variable.

It is interesting to see that the reported response from all the ASRMs, discussed above, can be written in a generalized form as:
\[
Z = \alpha_i X_i + (1 - \alpha_i) (X_i + Y_i) \tag{18}
\]
where \( X \) is the study variable with mean \( \mu_X \) and variance \( \sigma_X^2 \), \( Y \) is the scrambling variable with mean \( \mu_Y \) and variance \( \sigma_Y^2 \), and \( \alpha_i \) is Bernoulli random variable with \( E(\alpha_i) = A \). Then an unbiased estimator of population mean and its variance are given, respectively, by
\[
\hat{\mu}_X = \bar{Z} - (1 - A) \mu_Y \tag{19}
\]
and
\[
\text{Var}(\hat{\mu}_X) = \frac{1}{n} \left\{ \sigma_Y^2 + (1 - A) (\sigma_Y^2 + A \mu_Y^2) \right\} \tag{20}
\]
[28] proposed a new additive scrambled response model in which they use three statements. Each respondent is asked to rotate the spinner and report the answer to one of statements, randomly pointed by the spinner. The three statements are: (i) Report the true response \( X \) with probability, \( p_1 \), (ii) Report the
scrambled response $X S_1 + S_2$ with probability $p_2$ and (iii) Report the value of $S_3$ with probability $p_3$. The variables $S_1, S_2$ and $S_3$ are three scrambling variables. If we replace $p_1 = A, p_2 = 1 - A, S_1 = 1, S_2 = Y$ and $S_3 = 0$ then the generalized model given in (18) becomes a special case of the [28] model.

The rest of the article is arranged as follows. In section 2, we present proposed generalized model and show that most of the existing ASRMs are special cases of the proposed model and a list of some new models is also given. In section 3, we compare proposed model with existing models discussed in Section 1, and prove the superiority of our proposed models both analytically and numerically. In section 4, we measure the privacy protection of respondent. In the last section, we give the summary and concluding remarks of this article.

2. PROPOSED MODEL

Let we have a population $U = (u_1, u_2, ..., u_N)$ of size $N$ and a random sample of size $n$ is selected from the population. Respondents are selected by simple random sampling with replacement. Let $A$ be the proportion of sampled individuals asked to answer truthfully and the remaining proportion, $(1 - A)$, of the respondents is directed to report a scrambled response. The scrambling is done as follows. Each respondent (who randomly chooses to scramble his/her response) is provided a randomization device to generate the $G(>1)$ values of scrambling variable (following a pre-assigned distribution). For a given respondent, sample mean of scrambling variable is $\bar{y} = \frac{\sum_j^G y_j}{G}$. Then he/she is requested to add average $\bar{y}$ of $G$ values of scrambling variable to his/her true response. The randomization (scrambling) procedure is performed in such a way that the interviewer doesn’t know about the $G$ values of scrambling variable. Let $X$ be the study variable with mean $\mu_X$ and variance $\sigma_X^2$. Let $Y$ be the scrambling variable with mean $\mu_Y$ and variance $\sigma_Y^2$. Now, the $i^{th}$ reported response $Z_i$ is given by

$$Z_i = \begin{cases} X_i & \text{with probability } A \\ (X_i + \bar{y}_i) & \text{with probability } 1 - A \end{cases} \quad (21)$$

The expected response is given by

$$E(Z) = AE(X) + (1 - A)E(X + \bar{Y})$$

$$E(Z) = A\mu_X + (1 - A)(\mu_X + \mu_Y)$$

$$E(Z) = \mu_X + (1 - A)\mu_Y.$$ 

An estimator of the population mean $\mu_X$ may be suggested as:

$$\hat{\mu}_{XP} = \bar{Z} - (1 - A)\mu_Y. \quad (22)$$

The variance of the proposed estimator $\hat{\mu}_{XP}$ is given by

$$\text{Var}(\hat{\mu}_{XP}) = \frac{1}{n}\left\{\sigma_X^2 + (1 - A)\left(\frac{\sigma_Y^2}{G} + A\mu_Y^2\right)\right\}. \quad (23)$$

**Theorem 1:** $\hat{\mu}_{XP}$ is an unbiased estimator of population mean $\mu_X$.

*Proof:* Theorem 1 can easily be proved by using the fact that $E(\bar{Z}) = E(Z)$.

**Theorem 2:** The variance of the unbiased estimator $\hat{\mu}_{XP}$ is given by

$$\text{Var}(\hat{\mu}_{XP}) = \frac{1}{n}\left\{\sigma_X^2 + (1 - A)\left(\frac{\sigma_Y^2}{G} + A\mu_Y^2\right)\right\}.$$ 

*Proof:* We have

$$\text{Var}(\hat{\mu}_{XP}) = \frac{\sigma_Y^2}{n}.$$ 

To find $\sigma_Y^2$, consider
\[ \sigma_Z^2 = E(Z_i^2) - (E(Z_i))^2. \]

and

\[ E(Z_i^2) = \sigma_X^2 + \mu_Z^2 + (1 - A) \left\{ \frac{\sigma_X^2 + \mu_Y^2}{G} + \frac{G - 1}{G} \mu_Y^2 \right\} + 2(1 - A)\mu_X\mu_Y. \]

Hence, the Theorem 2 can easily be proved.

**Theorem 3:** The unbiased estimator of the variance of \( \hat{\mu}_{XP} \) is given by

\[ \text{Var}(\hat{\mu}_{XP}) = \frac{s_Z^2}{n}, \]

where \( s_Z^2 = \frac{1}{n-1} \sum_i (Z_i - \bar{Z})^2. \)

**Proof:** Theorem 3 can easily be proved by using the fact that \( E(s_Z^2) = \sigma_Z^2. \)

All the ASRM, discussed earlier, can be written as one stage model and, now, we show that all the models are special cases of the proposed model. Corresponding to [22] model, a new model may be proposed by setting \( A = 0 \) in Equation (21) as follows.

\[ Z_i = X + \bar{y}_i \] (24)

Now, the unbiased estimator of population mean and its variance are given, respectively, by

\[ \hat{\mu}_{XP1} = \bar{Z} \] (25)

and

\[ \text{Var}(\hat{\mu}_{XP1}) = \frac{(\sigma_X^2 + \sigma_Z^2)}{n}. \] (26)

Corresponding to the [23] model, distribution of reported response through proposed methodology is given by

\[ Z_i = \begin{cases} X & \text{with probability } P + (1 - P)(1 - W) \\ X + \bar{Y} & \text{with probability } W(1 - P) \end{cases} \] (27)

Using the above reported response, an unbiased estimator of population mean and its variance are now given, respectively, by

\[ \hat{\mu}_{XP2} = \frac{\theta_2 Z_1 - \theta_1 Z_2}{\theta_2 - \theta_1} \] (28)

and

\[ \text{Var}(\hat{\mu}_{XP2}) = \frac{1}{(\theta_2 - \theta_1)^2} \left\{ \theta_2^2 \frac{\sigma_X^2}{n_1} + \theta_1^2 \frac{\sigma_Y^2}{n_2} \right\}. \] (29)

where

\[ \sigma_Z^2 = \sigma_X^2 + \frac{\sigma_Y^2}{G} \left\{ (1 - P)W \right\} + \theta_1^2 \left\{ (1 - P)W \right\} [1 - (1 - P)W]. \]

Corresponding to the [24] model, the proposed structure generates the following distribution of reported response.

\[ Z_i = \begin{cases} X & \text{with probability } (1 - T)(1 - W) \\ X + \bar{Y} & \text{with probability } T + W(1 - T) \end{cases} \] (30)

An unbiased estimator of population mean and its variance are given, respectively, by

\[ \hat{\mu}_{XP3} = \frac{\theta_2 Z_1 - \theta_1 Z_2}{\theta_2 - \theta_1} \] (31)
and

$$\text{Var}(\hat{\mu}_{XP3}) = \frac{1}{(\theta_2 - \theta_1)^2} \left\{ \theta_2^2 \frac{\sigma_{Z_i}^2}{n_1} + \theta_1^2 \frac{\sigma_{Z_i}^2}{n_2} \right\},$$

(32)

where \(\sigma_{Z_i}^2 = \sigma_X^2 + \frac{\sigma_Y^2}{G}(T + (1 - T)W) + \theta_i^2[T + (1 - T)W][1 - T + (1 - T)W].\)

Following the proposal by [21], the expected response through the proposed model may be written as:

$$E(Z_i) = (1-T)X + T(S_iX + \bar{Y}).$$

(33)

Based on the above reported response, an unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XP4} = \frac{\theta_1 Z_2 - \theta_2 Z_1}{\theta_1 - \theta_2}$$

(34)

and

$$\text{Var}(\hat{\mu}_{XP4}) = \frac{1}{(\theta_1 - \theta_2)^2} \left\{ \theta_2^2 \frac{\sigma_{Z_i}^2}{n_2} + \theta_1^2 \frac{\sigma_{Z_i}^2}{n_1} \right\},$$

(35)

where \(\sigma_{Z_i}^2 = \sigma_X^2 + W \frac{\sigma_Y^2}{G}(\sigma_X^2 + \mu_X^2) + W\gamma_i^2 + W(1 - W)\theta_i^2\).

Similarly, corresponding to the [25] model, the proposed structure would generate the distribution of reported response given by

$$Z_i = \begin{cases} X & \text{with probability } P + (1 - P)(1 - W) \\ S_iX + \bar{Y} & \text{with probability } W(1 - P). \end{cases}$$

(36)

And, an unbiased estimator of population mean and its variance are given, respectively, by

$$\hat{\mu}_{XP5} = \frac{\theta_2 Z_1 - \theta_1 Z_2}{\theta_2 - \theta_1}$$

(37)

and

$$\text{Var}(\hat{\mu}_{XP5}) = \frac{1}{(\theta_1 - \theta_2)^2} \left\{ \theta_2^2 \frac{\sigma_{Z_i}^2}{n_1} + \theta_1^2 \frac{\sigma_{Z_i}^2}{n_2} \right\},$$

(38)

where \(\sigma_{Z_i}^2 = \sigma_X^2 + W(1 - P)\frac{\sigma_Y^2}{G}(\sigma_X^2 + \mu_X^2) + W(1 - P)(\gamma_i^2 + \theta_i^2) - W^2(1 - P)^2\theta_i^2.\)

It is interesting to note that for \(G = 1\), all the proposed models reduce to usual corresponding models.

If we replace \(p_1 = A, p_2 = 1 - A, S_1 = 1, S_2 = \bar{Y}\) and \(S_3 = 0\) in [28] model then our proposed generalized model becomes a special case of the [28] model.

### 3. EFFICIENCY COMPARISON

In this section, we carry out some analytical and numerical comparisons between the proposed models and some existing models.

Consider,

$$\text{Var}(\hat{\mu}_{XP}) < \text{Var}(\hat{\mu}_X)$$

Using Equations (20) and (23) in the above inequality, we get

$$\frac{1}{n} \left\{ \sigma_X^2 + (1 - A)(\frac{\sigma_Y^2}{G} + A\mu_X^2) \right\} < \frac{1}{n} \left\{ \sigma_X^2 + (1 - A)(\sigma_Y^2 + A\mu_X^2) \right\}$$
\[(1 - A)(\frac{\sigma^2}{G} + A\mu^2_Y) < (1 - A)(\sigma^2 + A\mu^2_Y)\]
\[\frac{\sigma^2}{G} < \sigma^2\]
\[G > 1\,.

which is always true since in the proposed structure we assumed \(G > 1\). Thus, the proposed methodology is better than the usual one where a single value of scrambling variable is used to scramble the response.

To know the extent of relative efficiency and prove the superiority of the proposed model a simulation study is conducted. The simulated variances for proposed and existing models (when \(G = 1\)) are calculated using random samples of size 100. Different values of \(A\) such as 0.1, 0.3, 0.5, 0.7 and 0.9 are considered.

The number of values of scrambling variable are considered as \(G = 1, 3\) and 5. The study variable \(X\) is assumed to be Normally distributed with mean 2 and variance 1. The scrambling variable \(Y\) is taken as Poisson random variable with mean \(\mu_Y = 3\). The simulated results based on 50000 iterations are showcased in Table 1, given below.

**Table 1. Comparison Table of \(\text{Var} (\hat{\mu}_X)\) and \(\text{Var} (\hat{\mu}_{XP})\)**

<table>
<thead>
<tr>
<th>(G)</th>
<th>(A)</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.04492325</td>
<td>0.04987604</td>
<td>0.0470403</td>
<td>0.03810669</td>
<td>0.02096469</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.02711408</td>
<td>0.03569062</td>
<td>0.03749994</td>
<td>0.03204375</td>
<td>0.01884359</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.02400066</td>
<td>0.03326997</td>
<td>0.03516758</td>
<td>0.03066921</td>
<td>0.01850494</td>
<td></td>
</tr>
</tbody>
</table>

The results in the Table 1 clearly show that the variance of the estimator based on the proposed generalized model is less than the variances of the estimator when \(G = 1\). The superiority of the proposed model can be easily proved by taking any of the models discussed in this paper and by taking any distribution of scrambling variable.

**4. PRIVACY PROTECTION**

In this section, we calculate the respondent’s degree of protection. That is, we try to find how much secure respondents feel when using the proposed model. We find the loss in privacy protection after using proposed model. A good statistical analysis or planning in randomized response is that through which we increase both the efficiency and privacy protection. We use a criterion proposed by [29]. To find degree of privacy protection, denoted by \(\Delta\), we take the expected value of the square of the difference of the response obtained by randomized response technique and the true response. Let \(R\) be the response of the respondent obtained by using any randomized response technique and \(X\) be the true response, then the measure of privacy protection defined by [29] is given by

\[\Delta = E(R - X)^2\,.

We use generalized model and proposed generalized model defined in Section 1 and 2 to check loss in privacy protection. The amount of privacy protection, \(\Delta_X\), for the generalized model is

\[\Delta_X = E(R - X)^2\]
\[\Delta_X = E(\alpha_iX_i + (1 - \alpha_i)(X_i + Y_i) - X)^2\]
\[\Delta_X = (1 - A)(\mu^2_Y + \sigma^2_Y)\]

Similarly, the amount of privacy protection, \(\Delta_{XP}\), for the proposed generalized model, explained in Section 2, is
$$\Delta_{XP} = E(R - X)^2$$
$$\Delta_{XP} = E(AX + (1 - A)(X + \bar{Y}) - X)^2$$
$$\Delta_{XP} = (1 - A)(\mu_Y^2 + \frac{\sigma_Y^2}{G})$$

Now, the loss in amount of privacy protection is

$$\Delta_{XP} = (1 - A)(\mu_Y^2 + \frac{\sigma_Y^2}{G}) - (1 - A)(\mu_Y^2 + \frac{\sigma_Y^2}{G})$$

$$\Delta_{XP} = (1 - A)\sigma_Y^2 \left(\frac{G - 1}{G}\right).$$  \hspace{1cm} (39)

To have a tangible idea about the loss in privacy, we calculate the percentage relative loss (PRL) defined as

$$\frac{\Delta_{X} - \Delta_{XP}}{\Delta_{X}} \times 100$$

for \( G = 2, 3 \), and different values of \( \mu_Y \). The results for PRL are presented in the Table 2.

**Table 2. The PRL for different values of \( \mu_Y \) and \( \sigma_Y^2 \)**

<table>
<thead>
<tr>
<th>( \sigma_Y^2 )</th>
<th>( \mu_Y = 2 )</th>
<th>( \mu_Y = 2.5 )</th>
<th>( \mu_Y = 3 )</th>
<th>( \mu_Y = 3.5 )</th>
<th>( \mu_Y = 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>27.77</td>
<td>25.00</td>
<td>22.72</td>
<td>20.83</td>
<td>19.23</td>
</tr>
<tr>
<td>3.0</td>
<td>30.00</td>
<td>27.27</td>
<td>25.00</td>
<td>23.07</td>
<td>21.42</td>
</tr>
<tr>
<td>3.5</td>
<td>31.81</td>
<td>29.16</td>
<td>26.92</td>
<td>25.00</td>
<td>23.37</td>
</tr>
<tr>
<td>4.0</td>
<td>33.33</td>
<td>30.30</td>
<td>28.57</td>
<td>26.00</td>
<td>25.00</td>
</tr>
<tr>
<td>4.5</td>
<td>34.61</td>
<td>32.14</td>
<td>30.00</td>
<td>28.12</td>
<td>26.47</td>
</tr>
<tr>
<td></td>
<td>37.03</td>
<td>33.33</td>
<td>30.30</td>
<td>27.77</td>
<td>25.64</td>
</tr>
<tr>
<td></td>
<td>40.00</td>
<td>36.36</td>
<td>33.33</td>
<td>30.76</td>
<td>28.57</td>
</tr>
<tr>
<td></td>
<td>42.42</td>
<td>38.88</td>
<td>35.89</td>
<td>33.33</td>
<td>31.11</td>
</tr>
<tr>
<td></td>
<td>44.44</td>
<td>41.02</td>
<td>38.09</td>
<td>35.55</td>
<td>33.33</td>
</tr>
<tr>
<td></td>
<td>46.15</td>
<td>42.85</td>
<td>40.00</td>
<td>37.00</td>
<td>35.29</td>
</tr>
</tbody>
</table>

From (39), we can see that as the value of \( G \) increases, the loss in privacy protection increases and for \( G = 1 \), both the models provide equal privacy protection. Also, from Table 2, it is observed that PRL decreases (increases) with the increase in \( \mu_Y \). As is obvious, for \( G > 1 \), our proposed model performs better than the other models in terms of efficiency. So, we check how much we gain in efficiency.

**Gain in efficiency**

$$\text{Gain in efficiency} = Var(\hat{\mu}_X) - Var(\hat{\mu}_{XP})$$

$$= \frac{1}{n} \{ \sigma_X^2 + (1 - A)(\sigma_Y^2 + A\mu_Y^2) \} - \frac{1}{n} \{ \sigma_X^2 + (1 - A)(\sigma_Y^2 + A\mu_Y^2) \}$$

$$= \frac{(1 - A)}{n} \left( \sigma_Y^2 + A\mu_Y^2 \right) - \frac{(1 - A)}{n} \left( \frac{\sigma_Y^2}{G} + A\mu_Y^2 \right)$$
Again, to have a clear picture of gain in efficiency, we calculate the percentage relative gain (PRG) defined as
\[
\text{PRG} = \left( \frac{\text{Var}(\hat{\mu}_x) - \text{Var}(\hat{\mu}_{\text{eq}})}{\text{Var}(\hat{\mu}_x)} \right) \times 100
\]
for \( G = 2, 3 \), and different values of \( \sigma^2_x \). The numerical values of PRG so obtained are arranged in the Table 3.

**Table 3. The PRG for \( \sigma^2_x \) and \( A \)**

<table>
<thead>
<tr>
<th>( \sigma^2_x )</th>
<th>( G = 2 )</th>
<th>( G = 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( A = 0.1 )</td>
<td>( A = 0.3 )</td>
</tr>
<tr>
<td>( \sigma^2_x = 1 )</td>
<td>32.80</td>
<td>30.12</td>
</tr>
<tr>
<td>( \sigma^2_x = 1.5 )</td>
<td>28.62</td>
<td>26.17</td>
</tr>
<tr>
<td>( \sigma^2_x = 2 )</td>
<td>25.40</td>
<td>23.15</td>
</tr>
<tr>
<td>( \sigma^2_x = 2.5 )</td>
<td>22.82</td>
<td>20.74</td>
</tr>
<tr>
<td>( \sigma^2_x = 2 )</td>
<td>20.71</td>
<td>18.79</td>
</tr>
</tbody>
</table>

From the results in Table 3, it is observed that PRG can be increased by increasing the value of \( G \). Similarly, the PRG is higher for smaller values of \( A \) and \( \sigma^2_x \).

From (39) and (40), it is clearly seen that we have loss in privacy protection and gain in efficiency when \( G > 1 \) is used. [29] proved that in every randomized response the efficiency decreases when amount of privacy protection increases and vice versa. They showed that increase in privacy and increase in efficiency are in conflict. If, we use the direct questioning method then we have the minimum variance (i.e. \( \frac{\sigma^2}{n} \)) but, in this case, we have to sacrifice the privacy of the respondent (amount of privacy protection is zero). In this article, we see that as the value of \( G \) increases we have smaller variance of the proposed estimator.

According to central limit theorem, as the value of \( G \) increases, the distribution of \( \bar{Y} \) approaches to normal distribution with mean \( \mu_Y \). As the value of \( G \) increases, we know that value of \( \bar{Y} \) would be closer to \( \mu_Y \) with higher probability. Thus, we would actually have good guess about the true response of a respondent. So, we definitely lose privacy of respondent for larger value of \( G \). Intuitively, we suggest that \( G \) should be less than 5 since respondents are less burdened as they are required to select a few values of scrambling variable. We also found this suggestion of fixing the number of values of scrambling variable at 2, 3 and 4 in [30]. Our intuition may also be supported by the observations from the Table 4, given below. From Table 4, it is obvious that most (75%) of the loss and gain is achieved at \( G = 4 \). Moving from \( G = 4 \) to \( G = 10 \) results in only 15% increase in “loss in privacy” and “gain in efficiency”. Obviously, when we move from \( G = 4 \) to onwards, only 25% increase in “loss in privacy” and “gain in efficiency” will be observed. Thus, to keep the proposed model more parsimonious setting \( G < 5 \) seems more practicable.
Table 4. Percentage of “loss in privacy” and “gain in efficiency” for different values of $G$ 

<table>
<thead>
<tr>
<th>$G$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss and Gain</td>
<td>50%</td>
<td>67.66%</td>
<td>75%</td>
<td>80%</td>
<td>83.33%</td>
<td>85.71%</td>
<td>87.5%</td>
<td>88.88%</td>
<td>90%</td>
</tr>
</tbody>
</table>

5. SUMMARY AND CONCLUSION

The idea of increasing the protection of respondent’s privacy through scrambling of the response is very common today. The researchers use randomized response devices to perturb the value of sensitive variable by adding, subtracting or multiplying the value of scrambling variable whose distribution is known. In this article, we gave the respondent a new device which requires adding average of $G$ random values of a scrambling variable to the value of sensitive variable. The main advantage behind this idea is that it increases the respondent’s cooperation without increasing the sampling cost and losing efficiency. As special cases, different additive scrambled response models are discussed. A complete theoretical comparison of existing models and the proposed model is presented and superiority of the proposed model is established. A simulation based study is also conducted to know the extent of better performance of the proposed method. We also checked the respondent privacy protection whether it remains same or not. We conclude that for $G > 1$, we have to sacrifice same percentage of amount of respondent privacy as the increase in efficiency.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES


