

Calibration-Based Improvements in Estimating the Mean of a Stratified Population with Scrambled Response on Second Call under Non-Response

Manoj K. Chaudhary¹ and Tulika Dutta²

^{1,2}Department of Statistics, Institute of Science,
Banaras Hindu University, Varanasi, INDIA

¹Email: mk15@bhu.ac.in

²Email: duttatulika7@gmail.com

ABSTRACT

This paper presents some salient features on the calibration-based improvements in estimating the mean of a stratified population with scrambled response on second call under non-response. We have proposed a new calibration estimator of the population mean under scrambled randomized response method utilizing the information on a single auxiliary variable. The proposed calibration estimator has been pointed out under some existing scrambled response models. We have derived the expression for the MSE of the proposed calibration estimator. We have also discussed the privacy protection measure to get the trade-off between efficiency and privacy protection. A simulation study has been carried out to depict the performance of the proposed calibration estimator.

Keywords: Stratified random sampling, calibration approach, population mean, scrambled response, privacy protection.

Mathematics Subject Classification: 62D05, 62D99

1. INTRODUCTION

Non-response problem is a big issue in almost all types of survey and it occurs when people don't give the required information whenever they are contacted through mail questionnaire or telephonic interview. The occurrence of non-response on surveys depends on the type of information to be required. The non-response occurs due to unavailability of people, lack of interest or unable to understand the questions in the cases where the variable which we want to study is related to the general issues. But, in some cases, people hesitate to provide the true information when questions asked to them are based on the sensitive characters or social undesirability e.g., gambling, illegal

income, tax evasion, drug addiction, etc. To overcome the problem of non-response, Hansen and Hurwitz (1946) suggested a procedure of estimating the finite population mean by sub-sampling the non-respondents with the assumption that the people who are selected for second call, provide the true information. Adopting the technique of sub-sampling of non-respondents, several authors have suggested a number of estimators utilizing the auxiliary information which may or may not suffer from the non-response. Khare and Pandey (2000) proposed a class of estimators for ratio of two population means utilizing the auxiliary information in the presence of non response. Chaudhary et al. (2012) suggested a class of estimators for population mean using auxiliary information under non-response. Shabbir and Khan (2013) suggested the estimator of finite population mean using two auxiliary variables in two-phase sampling under non-response.

But, if the required information is of sensitive nature, the people who refused to provide information on first call also refused to answer on second call once they are selected in sub-sample. In such a situation, the Hansen and Hurwitz (1946) technique is not useful. In literature, many authors have dealt with the problem of non-response due to sensitiveness of the character under study and introduced the estimators to maintain the confidentiality of the responses provided by the respondents. Some randomized response methods are generally used to maintain the confidentiality of the response. First, Warner (1965) applied this method to estimate the proportion of population having a sensitive attribute which requires a 'yes' or 'no' response or selects response from a set of nominal categories. Using Warner (1965) method of randomization, many authors such as Fox and Tracy (1986), Mangat and Singh (1990), Shabbir and Gupta (2005) and Diana and Perri (2009, 2010) have introduced estimators for which the responses given under sensitive character are of qualitative in nature. Greenberg et al. (1969) and Erikson (1973) have dealt with the situations in which the responses of the sensitive questions are of quantitative nature. Some authors have introduced the idea of scrambled randomized response (SRR) method to deal with the problems of such situations. In SRR method, the respondents multiply their sensitive answers by random numbers generated through known distribution and provide scrambled response to the interviewer who does not have an idea about the random numbers. The estimators under SRR method are less efficient but high in privacy protection.

To overcome the efficiency loss, one can utilize the auxiliary information. It is well known fact that the use of auxiliary information increases the precision of the estimates. Several authors have introduced the estimators using auxiliary information in SRR method. Diana et al. (2014) have proposed Hansen and Hurwitz-type estimator with scrambled response on second call. Ahmed et al. (2016) have suggested Hansen and Hurwitz-type estimator with scrambled response on second call in stratified random sampling. In the subsequent sections, we have proposed a calibration estimator under SRR method to estimate the finite population mean utilizing the information on a single auxiliary variable. The objective of the present study is to develop the estimator which improves the efficiency and maintains the privacy too. To improve the efficiency of the estimator, the information on an auxiliary

variable has been used and some SRR methods have been used to maintain the confidentiality of the sensitive information.

2. SAMPLING DESIGN, NOTATIONS AND TERMINOLOGY

Consider a population of N distinct and identifiable units which is divided into L strata. The h^{th} stratum consists of N_h units ($h=1,2,...,L$) such that $\sum_{h=1}^L N_h = N$. Let Y and X be the study and auxiliary variables with respective population means $\bar{Y} (= \sum_{h=1}^L p_h \bar{Y}_h)$ and $\bar{X} (= \sum_{h=1}^L p_h \bar{X}_h)$. Here, $\bar{Y}_h$ and $\bar{X}_h$ are respectively the stratum means under study and auxiliary variables for the h^{th} stratum; $p_h (= \frac{N_h}{N})$ is the weight for the h^{th} stratum. Let y_{hi} and x_{hi} be the observations on the i^{th} unit in the h^{th} stratum under study and auxiliary variables respectively ($i=1,2,...,N_h$). It is assumed that the non-response occurs on only study variable due to sensitiveness of the character under study. A sample of size n_h is selected from the h^{th} stratum using simple random sampling without replacement (SRSWOR) scheme out of which only n_{h1} units respond and rest n_{h2} units don't respond on the first call under study variable. Now, we select a sub-sample of size $r_{h2} (= \frac{n_{h2}}{k_h}; k_h > 1)$ from n_{h2} non-responding units but due to sensitiveness of the character under study, the units selected in the sub-sample again do not respond on the second call. Due to this reason, respondents are now allowed to scramble their responses. The aim behind that is to ensure the confidentiality of the responses given by the respondents so that they would provide truthful information. Here, we have used linear model given by Diana et al. (2014). Let Z be the scrambled response which is based on two independent scrambled random variables U and V unrelated with Y and having means μ_U , μ_V and variances σ_U^2 , σ_V^2 respectively. Suppose Y_h , Z_h , U_h and V_h are respectively the variables under Y , Z , U and V for the h^{th} stratum.

Now, we can write the randomized linear model as

$$Z_h = U_h Y_h + V_h \quad (1)$$

with mean $E_R(Z_h) = \mu_{U_h} Y_h + \mu_{V_h}$ and variance $V_R(Z_h) = \sigma_{U_h}^2 Y_h^2 + \sigma_{V_h}^2$

where μ_{U_h} , μ_{V_h} and $\sigma_{U_h}^2$, $\sigma_{V_h}^2$ are respectively the means and variances for the variables U_h and V_h .

It is assumed that the values generated by the respondents from the scrambling distributions of U_h and V_h are not known to the interviewer and hence their privacy is protected. Let $\hat{y}_{hi}$ and z_{hi} respectively be the transformed randomized response and scrambled response of the i^{th} unit in the h^{th} stratum, i.e.

$$\hat{y}_{hi} = \frac{z_{hi} - \mu_{V_h}}{\mu_{U_h}} \quad (2)$$

such that

$$E_R(\hat{y}_{hi}) = y_{hi} \quad (3)$$

Thus, the expectation under randomization coincides with the true response y_{hi} in the h^{th} stratum.

The variance of $\hat{y}_{hi}$ under randomization is given as

$$V_R(\hat{y}_{hi}) = \frac{\sigma_{U_h}^2 y_{hi}^2 + \sigma_{V_h}^2}{\mu_{U_h}^2} = \phi_{hi} \quad (4)$$

Here, we have also used two other models to get some idea about efficiency and privacy protection. Let,

M_1 = Saha Mixed Model (2007) and M_2 = Pollock and Bek Additive Model (1976).

Under the model M_1 , Z_h is defined as

$$Z_h = U_h(Y_h + V_h) \quad (5)$$

Under the model M_2 , Z_h is given as

$$Z_h = Y_h + V_h \quad (6)$$

We can easily see that the expectation of $\hat{y}_{hi}$ under the models M_1 and M_2 coincides with the true response y_{hi} in the h^{th} stratum. Thus, the expressions for variance of $\hat{y}_{hi}$ under the models M_1 and M_2 are respectively given as

$$V_R(\hat{y}_{hi})_{M_1} = \frac{\sigma_{Uh}^2 y_{hi}^2 + \sigma_{Uh}^2 \sigma_{Vh}^2 + \sigma_{Uh}^2 \mu_{Vh}^2 + \sigma_{Vh}^2 \mu_{Uh}^2 + 2y_{hi} \sigma_{Uh}^2 \mu_{Vh}}{\mu_{Uh}^2} = \varphi_{hi(M_1)} \quad \text{and}$$

$$V_R(\hat{y}_{hi})_{M_2} = \sigma_{Vh}^2 = \varphi_{hi(M_2)}.$$

The Hansen and Hurwitz (1946) estimator of the population mean $\bar{Y}$ under stratified random sampling with scrambled response on second call is given as

$$\hat{y}_{st} = \sum_{h=1}^L p_h \hat{y}_h \quad (7)$$

$$\text{where } \hat{y}_h = w_{h1} \bar{y}_{h1} + w_{h2} \hat{y}_{h(2)}, \quad w_{h1} = \frac{n_{h1}}{n_h}, \quad w_{h2} = \frac{n_{h2}}{n_h}, \quad \bar{y}_{h1} = \frac{1}{n_{h1}} \sum_i^{n_{h1}} y_{hi} \quad \text{and} \quad \hat{y}_{h(2)} = \frac{1}{r_{h2}} \sum_i^{r_{h2}} \hat{y}_{hi}.$$

Since, $\hat{y}_{st}$ is an unbiased estimator of the population mean $\bar{Y}$, hence expression for the variance of the estimator $\hat{y}_{st}$ is given by

$$V(\hat{y}_{st}) = \sum_{h=1}^L p_h^2 \left[\left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{Yh}^2 + W_{h2} \frac{(k_h - 1)}{n_h} S_{Yh2}^2 + \frac{k_h}{n_h N_h} \sum_i^{N_{h2}} \varphi_{hi} \right] \quad (8)$$

where $\frac{1}{N_{h2}} \sum_i^{N_{h2}} \varphi_{hi} = \frac{\sigma_{Uh}^2 \mu_{Yh2} + \sigma_{Vh}^2}{\mu_{Uh}^2}$, $\mu_{Yh2} = S_{Yh2}^2 + \bar{Y}_{h2}^2$, $S_{Yh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2$, $S_{Yh2}^2 = \frac{1}{N_{h2} - 1} \sum_i^{N_{h2}} (y_{hi} - \bar{Y}_{h2})^2$, $\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}$ and $\bar{Y}_{h2} = \frac{1}{N_{h2}} \sum_i^{N_{h2}} y_{hi}$. N_{h2} is the size of the non-response group in the h^{th} stratum.

3. PROPOSED CALIBRATION ESTIMATOR

The Hansen and Hurwitz (1946) estimator of the population total $T (= \sum_{h=1}^L \sum_{i=1}^{N_h} y_{hi})$ under stratified random sampling with scrambled response on second call is given as

$$\hat{y}_{st} = \sum_{h=1}^L \left[\sum_i^{n_{h1}} d_{hi} y_{hi} + \sum_i^{r_{h2}} d_{hi}^* \hat{y}_{hi} \right] \quad (9)$$

Here, the inclusion probabilities are defined as

$$\Pi_{hi} = \frac{n_h}{N_h} \text{ and } \Pi_{hi|n_{h2}} = \frac{r_{h2}}{n_{h2}}.$$

Using these values of the inclusion probabilities, the design weights become

$$d_{hi} = \frac{1}{\Pi_{hi}} = \frac{N_h}{n_h} \text{ and } d_{hi}^* = \frac{1}{\Pi_{hi}\Pi_{hi|n_{h2}}} = \frac{N_h}{n_h} \frac{n_{h2}}{r_{h2}}.$$

We now propose a new calibration estimator of the population total T in stratified random sampling with scrambled response on second call under non-response as follows:

$$\hat{y}_t = \sum_{h=1}^L \left[\sum_i^{n_{h1}} d_{hi} y_{hi} + \sum_i^{r_{h2}} \Omega_{hi} \hat{y}_{hi} \right] \quad (10)$$

where Ω_{hi} is the calibrated weight.

The calibrated weight Ω_{hi} is so chosen that it minimizes the chi-square type distance

$$\sum_i^{r_{h2}} \frac{(\Omega_{hi} - d_{hi}^*)^2}{q_{hi} d_{hi}^*}$$

subject to the calibration constraint

$$\sum_i^{r_{h2}} \Omega_{hi} x_{hi} = \sum_i^{n_{h2}} d_{hi} x_{hi} \quad (11)$$

where q_{hi} is the tuning parameter.

We now define the Lagrange function as

$$\Delta_1 = \sum_i^{r_{h2}} \frac{(\Omega_{hi} - d_{hi}^*)^2}{q_{hi} d_{hi}^*} - 2\lambda_1 \left[\sum_i^{r_{h2}} \Omega_{hi} x_{hi} - \sum_i^{n_{h2}} d_{hi} x_{hi} \right]$$

where λ_1 is the Lagrange multiplier.

Differentiating Δ_1 with respect to Ω_{hi} and equating the derivative to zero, we get

$$\Omega_{hi} = d_{hi}^* + \lambda_1 q_{hi} d_{hi}^* x_{hi} \quad (12)$$

Putting the value of Ω_{hi} from equation (12) into the equation (11), the value of λ_1 becomes

$$\lambda_1 = \frac{\sum_i^{n_{h2}} d_{hi} x_{hi} - \sum_i^{r_{h2}} d_{hi}^* x_{hi}}{\sum_i^{r_{h2}} q_{hi} d_{hi}^* x_{hi}^2}.$$

Substituting the value of λ_1 into the equation (12) and then assuming $q_{hi} = \frac{1}{x_{hi}}$, we get

$$\Omega_{hi} = d_{hi}^* \frac{\sum_i^{n_{h2}} d_{hi} x_{hi}}{\sum_i^{r_{h2}} d_{hi}^* x_{hi}}.$$

Putting the value of Ω_{hi} into the equation (10), we get

$$\hat{y}_t = \sum_{h=1}^L \left[\sum_i^{n_{h1}} d_{hi} y_{hi} + \frac{\sum_i^{n_{h2}} d_{hi} x_{hi}}{\sum_i^{r_{h2}} d_{hi}^* x_{hi}} \sum_i^{r_{h2}} d_{hi}^* \hat{y}_{hi} \right]$$

Substituting the values of design weights d_{hi} and d_{hi}^* , the calibration estimator of the population total T is given as

$$\hat{y}_t = \sum_{h=1}^L N_h [w_{h1} \bar{y}_{h1} + w_{h2} \frac{\hat{\bar{y}}_{h(2)}}{\bar{x}_{h(2)}} \bar{x}_{h2}]$$

where $\bar{x}_{h2} = \frac{1}{n_{h2}} \sum_i^{n_{h2}} x_{hi}$ and $\bar{x}_{h(2)} = \frac{1}{r_{h2}} \sum_i^{r_{h2}} x_{hi}$.

Thus, the calibration estimator for the population mean $\bar{Y}$ in stratified random sampling with scrambled response on second call under non-response is suggested as

$$\hat{\bar{y}}_{HH(cal)} = \sum_{h=1}^L p_h^* [w_{h1} \bar{y}_{h1} + w_{h2} \frac{\hat{\bar{y}}_{h(2)}}{\bar{x}_{h(2)}} \bar{x}_{h2}] \quad (13)$$

Here, p_h^* is the calibrated stratum weight such that it minimizes the chi-square type distance

$$\sum_{h=1}^L \frac{(p_h^* - p_h)^2}{q_h p_h} \quad (14)$$

subject to the calibration constraint

$$\sum_{h=1}^L p_h^* \bar{x}_h = \sum_{h=1}^L p_h \bar{X}_h \quad (15)$$

where $\bar{x}_h = w_{h1} \bar{x}_{h1} + w_{h2} \bar{x}_{h(2)}$ and $\bar{x}_{h1} = \frac{1}{n_{h1}} \sum_i^{n_{h1}} x_{hi}$. q_h is another tuning parameter.

The Lagrange function can be written as

$$\Delta_2 = \sum_{h=1}^L \frac{(p_h^* - p_h)^2}{q_h p_h} - 2\lambda_2 \left(\sum_{h=1}^L p_h^* \bar{x}_h - \sum_{h=1}^L p_h \bar{X}_h \right)$$

where λ_2 is the Lagrange multiplier.

Let us differentiate Δ_2 with respect to p_h^* and equate the derivative to zero, i.e.

$$\begin{aligned} \frac{\partial \Delta_2}{\partial p_h^*} &= \frac{2(p_h^* - p_h)}{q_h p_h} - 2\lambda_2 \bar{x}_h = 0 \\ \Rightarrow p_h^* &= p_h + \lambda_2 q_h p_h \bar{x}_h \end{aligned} \quad (16)$$

Putting the value of p_h^* from equation (16) into the equation (15), we get

$$\lambda_2 = \frac{\sum_{h=1}^L p_h \bar{X}_h - \sum_{h=1}^L p_h \bar{x}_h}{\sum_{h=1}^L q_h p_h \bar{x}_h}.$$

Substituting the value of λ_2 into equation (16) and then assuming $q_h = \frac{1}{x_h}$, the optimum stratum

weight p_h^* leads to

$$p_h^* = \frac{\sum_{h=1}^L p_h \bar{X}_h}{\sum_{h=1}^L p_h \bar{x}_h} p_h \quad (17)$$

Putting the value of p_h^* from equation (17) into the equation (13), the proposed calibration estimator becomes

$$\hat{y}_{HH(cal)} = \frac{\sum_{h=1}^L p_h [w_{h1} \bar{y}_{h1} + w_{h2} \frac{\hat{y}_{h(2)}}{\bar{x}_{h(2)}} \bar{x}_{h2}]}{\sum_{h=1}^L p_h [w_{h1} \bar{x}_{h1} + w_{h2} \bar{x}_{h(2)}]} \bar{X} \quad (18)$$

4. PROPERTIES OF THE PROPOSED CALIBRATION ESTIMATOR

Let $\bar{y}_{st} = \sum_{h=1}^L p_h [w_{h1} \bar{y}_{h1} + w_{h2} \frac{\hat{y}_{h(2)}}{x_{h(2)}} \bar{x}_{h2}]$ and $\bar{x}_{st} = \sum_{h=1}^L p_h [w_{h1} \bar{x}_{h1} + w_{h2} \bar{x}_{h(2)}]$. Thus, the proposed calibration estimator reduces to

$$\hat{y}_{HH(cal)} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X} \quad (19)$$

To obtain the mean square error (MSE) of the proposed calibration estimator $\hat{y}_{HH(cal)}$, we use the theory of large sample approximations.

Let us assume $\bar{y}_{st} = \bar{Y}(1 + e_0)$ and $\bar{x}_{st} = \bar{X}(1 + e_1)$ such that

$$E(e_0) = \frac{Bias(\bar{y}_{st})}{\bar{Y}}, \quad E(e_1) = 0, \quad E(e_0^2) = \frac{V(\bar{y}_{st})}{\bar{Y}^2} + \frac{[Bias(\bar{y}_{st})]^2}{\bar{Y}^2}, \quad E(e_1^2) = V(e_1) = \frac{V(\bar{x}_{st})}{\bar{X}^2} \quad \text{and}$$

$$E(e_0 e_1) = \frac{Cov(\bar{y}_{st}, \bar{x}_{st})}{\bar{Y} \bar{X}}.$$

Furthermore, we have

$$Bias(\bar{y}_{st}) = \sum_{h=1}^L p_h \frac{(k_h - 1)}{n_h} \left[\left(\frac{\bar{Y}_{h2} S_{Xh2}^2}{\bar{X}_{h2}^2} - \frac{S_{XYh2}}{\bar{X}_{h2}} \right) + \frac{\bar{Y}_{h2} S_{Xh2}^2}{\bar{X}_{h2}^2} \left(\frac{1}{n_{h2}} - \frac{1}{N_{h2}} \right) (3C_{Xh2}^2 - 2C_{Xh2} \lambda_{03h2} - 2\rho_{YXh2} C_{Xh2} C_{Yh2} + C_{Yh2} \lambda_{12h2}) - \frac{S_{XYh2}}{\bar{X}_{h2}} \left(\frac{1}{n_{h2}} - \frac{1}{N_{h2}} \right) (C_{Xh2}^2 - C_{Xh2} \frac{\lambda_{12h2}}{\rho_{YXh2}}) \right] \quad (20)$$

$$V(\bar{y}_{st}) = \sum_{h=1}^L p_h^2 \left[\left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{Yh}^2 + W_{h2} \frac{(k_h - 1)}{n_h} (S_{Yh2}^2 + R_h^2 S_{Xh2}^2 - 2R_h S_{XYh2}) + \frac{k_h}{n_h N_h} \sum_i^{N_{h2}} \varphi_{hi} \right] \quad (21)$$

$$V(\bar{x}_{st}) = \sum_{h=1}^L p_h^2 \left[\left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{Xh}^2 + W_{h2} \frac{(k_h - 1)}{n_h} S_{Xh2}^2 \right] \quad (22)$$

$$Cov(\bar{y}_{st}, \bar{x}_{st}) = \sum_{h=1}^L p_h^2 \left[\left(\frac{1}{n_h} - \frac{1}{N_h} \right) \rho_{YXh} S_{Yh} S_{Xh} + W_{h2} \frac{(k_h - 1)}{n_h} (\rho_{YXh2} S_{Yh2} S_{Xh2} - R_h S_{Xh2}^2) \right] \quad (23)$$

where

$$S_{Xh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (x_{hi} - \bar{X}_h)^2, \quad S_{Xh2}^2 = \frac{1}{N_{h2} - 1} \sum_{i=1}^{N_{h2}} (x_{hi} - \bar{X}_{h2})^2,$$

$$S_{XYh2} = \frac{1}{N_{h2} - 1} \sum_{i=1}^{N_{h2}} (x_{hi} - \bar{X}_{h2})(y_{hi} - \bar{Y}_{h2}), \quad \bar{X}_{h2} = \frac{1}{N_{h2}} \sum_{i=1}^{N_{h2}} x_{hi}, \quad C_{Xh2} = \frac{S_{Xh2}}{\bar{X}_{h2}}, \quad C_{Yh2} = \frac{S_{Yh2}}{\bar{Y}_{h2}},$$

$$\lambda_{rsh2} = \frac{\mu_{rsh2}}{\mu_{20h2}^{r/2} \mu_{02h2}^{s/2}}, \quad \mu_{rsh2} = \frac{1}{(N_{h2} - 1)} \sum_{i=1}^{N_{h2}} (y_i - \bar{Y}_{h2})^r (x_i - \bar{X}_{h2})^s, \quad W_{h2} = \frac{N_{h2}}{N_h} \quad \text{and} \quad R_h = \frac{\bar{Y}_{h2}}{\bar{X}_{h2}}.$$

ρ_{YXh} and ρ_{YXh2} are the correlation coefficients between Y and X for the entire and non-response groups respectively in the h^{th} stratum. N_{h2} is the number of non-responding units in the h^{th} stratum.

Now expressing the equation (19) in terms of e_0 and e_1 and then neglecting the higher order terms of e_0 and e_1 greater than two, we get

$$\hat{y}_{HH(cal)} - \bar{Y} = \bar{Y}(e_0 - e_1) \quad (24)$$

Squaring both sides of equation (24) and then taking expectation, we get

$$E(\hat{y}_{HH(cal)} - \bar{Y})^2 = \bar{Y}^2 [E(e_0^2) + E(e_1^2) - 2E(e_0 e_1)]$$

$$\Rightarrow MSE(\hat{y}_{HH(cal)}) = \bar{Y}^2 \left[\frac{V(\bar{y}_{st}) + (Bias(\bar{y}_{st}))^2}{\bar{Y}^2} + \frac{V(\bar{x}_{st})}{\bar{X}^2} - 2 \frac{Cov(\bar{y}_{st}, \bar{x}_{st})}{\bar{Y}\bar{X}} \right]$$

$$= V(\bar{y}_{st}) + (Bias(\bar{y}_{st}))^2 + R^2 V(\bar{x}_{st}) - 2RCov(\bar{y}_{st}, \bar{x}_{st}) \quad (25)$$

where $R = \frac{\bar{Y}}{\bar{X}}$.

From equation (25), we can obtain the expression for the MSE of the proposed calibration estimator $\hat{y}_{HH(cal)}$ up to the first order of approximation as

$$\begin{aligned}
 MSE(\hat{y}_{HH(cal)}) = & \sum_{h=1}^L p_h^2 \left[\left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{Yh}^2 + W_{h2} \frac{(k_h - 1)}{n_h} (S_{Yh2}^2 + R_h^2 S_{Xh2}^2 - 2R_h S_{XYh2}) + \frac{k_h}{n_h N_h} \sum_i^{N_{h2}} \phi_{hi} \right] \\
 & + \left[\sum_{h=1}^L p_h \frac{(k_h - 1)}{n_h} \left\{ \left(\frac{\bar{Y}_{h2} S_{Xh2}^2}{\bar{X}_{h2}^2} - \frac{S_{XYh2}}{\bar{X}_{h2}} \right) + \right. \right. \\
 & \left. \left. \frac{\bar{Y}_{h2} S_{Xh2}^2}{\bar{X}_{h2}^2} \left(\frac{1}{n_{h2}} - \frac{1}{N_{h2}} \right) (3C_{Xh2}^2 - 2C_{Xh2} \lambda_{03h2} - 2\rho_{YXh2} C_{Xh2} C_{Yh2} + C_{Yh2} \lambda_{12h2}) \right. \right. \\
 & \left. \left. - \frac{S_{XYh2}}{\bar{X}_{h2}} \left(\frac{1}{n_{h2}} - \frac{1}{N_{h2}} \right) (C_{Xh2}^2 - C_{Xh2} \frac{\lambda_{12h2}}{\rho_{YXh2}}) \right\}^2 + R^2 \sum_{h=1}^L p_h^2 \left[\left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{Xh}^2 + W_{h2} \frac{(k_h - 1)}{n_h} S_{Xh2}^2 \right] \right. \\
 & \left. - 2R \sum_{h=1}^L p_h^2 \left[\left(\frac{1}{n_h} - \frac{1}{N_h} \right) \rho_{YXh} S_{Yh} S_{Xh} + W_{h2} \frac{(k_h - 1)}{n_h} (\rho_{YXh2} S_{Yh2} S_{Xh2} - R_h S_{Xh2}^2) \right] \right] \quad (26)
 \end{aligned}$$

Remark 1: The equation (26) depicts the expression for the MSE of the proposed calibration estimator under the scrambled model given by Diana et al. (2014). It is to be noted that one can also obtain the expressions for the MSE of the proposed calibration estimator under the scrambled response models M_1 and M_2 in the similar manner. The MSE under scrambled response models M_1 and M_2 can be obtained by replacing ϕ_{hi} with $\phi_{hi(M_1)}$ and $\phi_{hi(M_2)}$ respectively.

5. SIMULATION STUDY

To get some idea about the efficiency and privacy protection, we have generated an artificial data set which describes a population of six strata with respective sizes 600, 500, 300, 800, 400 and 700. Here, we have generated the observations on the study variable Y for each stratum from the Normal distribution with certain mean and standard deviation (S.D.) amid 10%, 20% and 30% weights of the non-response group. We have also generated the observations on the auxiliary variable X for each stratum by assuming certain correlation with the study variable Y [See Reddy et al. (2010)]. To scramble the response, we have generated the values on two scrambled variables U and V from the Uniform distribution $U(0,1)$ which are not known to the interviewer. The assumed mean, standard deviation and correlation coefficient are given in Table 1.

Table 1: Particulars of population

Stratum No. (h)	Stratum Population for Study Variable Y	Stratum Size (N_h)	Correlation Coefficient (ρ_{YXh})
1	$N(\text{Mean} = 10, S.D. = 6.5)$	600	0.9009
2	$N(\text{Mean} = 15, S.D. = 6.7)$	500	0.9190
3	$N(\text{Mean} = 20, S.D. = 6.9)$	300	0.8706
4	$N(\text{Mean} = 25, S.D. = 7.1)$	800	0.9045
5	$N(\text{Mean} = 30, S.D. = 7.3)$	400	0.8969
6	$N(\text{Mean} = 35, S.D. = 7.5)$	700	0.9184

Here $N=3300$, $n=990$. The sample size for each stratum has been determined using proportional allocation. Now, we select the sample from each stratum using SRSWOR scheme and find the estimates of the population mean $\bar{Y}$ using the estimators $\hat{y}_{st}$ and $\hat{y}_{HH(cal)}$. The process of selecting the sample from each stratum and computing the estimates has been replicated 5000 times. Finally, we have computed the approximate variance/MSE (AVAR/AMSE) of the estimators $\hat{y}_{st}$ and $\hat{y}_{HH(cal)}$ using the following formulae:

$$AVAR(\hat{y}_{st}) = \frac{1}{5000} \sum_{l=1}^{5000} (\hat{y}_{stl} - \bar{Y})^2 ; l = 1, 2, \dots, 5000$$

$$AMSE(\hat{y}_{HH(cal)}) = \frac{1}{5000} \sum_{l=1}^{5000} (\hat{y}_{HH(cal)l} - \bar{Y})^2 ; l = 1, 2, \dots, 5000$$

Table 2 shows the AVAR of $\hat{y}_{st}$, AMSE of $\hat{y}_{HH(cal)}$ and percentage relative efficiency (PRE) of $\hat{y}_{HH(cal)}$ with respect to $\hat{y}_{st}$ under the scrambled response model given by Diana et al. (2014).

Table 2: AVAR of $\hat{y}_{st}$, AMSE of $\hat{y}_{HH(cal)}$ and PRE of $\hat{y}_{HH(cal)}$

$W_{h2} \forall h$	$k_h \forall h$	$AVAR(\hat{y}_{st})$	$AMSE(\hat{y}_{HH(cal)})$	$PRE(\hat{y}_{HH(cal)})$
0.1	1.5	0.04319375	0.01164378	370.9598
	2.0	0.04641665	0.01357645	341.8910
	2.5	0.04719780	0.01585650	297.6559
0.2	1.5	0.04686084	0.01596726	293.4808
	2.0	0.05368302	0.02068378	259.5416
	2.5	0.06114451	0.02445997	249.9779
0.3	1.5	0.05429165	0.01966932	276.0220
	2.0	0.06311002	0.02711705	232.7318
	2.5	0.07465310	0.03480076	214.5158

Table 3 depicts the AVAR of $\hat{y}_{st}$, AMSE of $\hat{y}_{HH(cal)}$ and PRE of $\hat{y}_{HH(cal)}$ with respect to $\hat{y}_{st}$ under the model M_1 i.e., Saha (2007) mixed model.

Table 3: AVAR of $\hat{y}_{st}$, AMSE of $\hat{y}_{HH(cal)}$ and PRE of $\hat{y}_{HH(cal)}$ under model M_1

$W_{h2} \forall h$	$k_h \forall h$	$AVAR(\hat{y}_{stM_1})$	$AMSE(\hat{y}_{HH(cal)M_1})$	$PRE(\hat{y}_{HH(cal)M_1})$
0.1	1.5	0.04318308	0.01164470	370.8390
	2.0	0.04642994	0.01358414	341.7952
	2.5	0.04720903	0.01586551	297.5577
0.2	1.5	0.04689387	0.01598549	293.3528
	2.0	0.05369946	0.02070476	259.3581
	2.5	0.06121175	0.02449537	249.8911
0.3	1.5	0.05432556	0.01969035	275.8994
	2.0	0.06315842	0.02714659	232.6569
	2.5	0.07466806	0.03483048	214.3756

Table 4 represents the AVAR of $\hat{y}_{st}$, AMSE of $\hat{y}_{HH(cal)}$ and PRE of $\hat{y}_{HH(cal)}$ with respect to $\hat{y}_{st}$ under the model M_2 , i.e., Pollock and Bek (1976) model.

Table 4: AVAR of $\hat{y}_{st}$, AMSE of $\hat{y}_{HH(cal)}$ and PRE of $\hat{y}_{HH(cal)}$ under model M_2

$W_{h2} \forall h$	$k_h \forall h$	$AVAR(\hat{y}_{stM_2})$	$AMSE(\hat{y}_{HH(cal)M_2})$	$PRE(\hat{y}_{HH(cal)M_2})$
0.1	1.5	0.04043967	0.008994237	449.6176
	2.0	0.04406120	0.01055285	417.5287
	2.5	0.04361727	0.01184644	368.1887
0.2	1.5	0.04128193	0.01051225	392.7029
	2.0	0.04667184	0.01346721	346.5591
	2.5	0.05245767	0.01588660	330.2006
0.3	1.5	0.04640356	0.01167502	397.4602
	2.0	0.05334680	0.01700002	313.8043
	2.5	0.06124222	0.02157709	283.8298

Table 5 depicts the normalized privacy protection measure for each stratum, τ_h given by Diana et al. (2013) under the Diana et al. (2014) linear model, Saha (2007) mixed model and Pollock and Bek (1976) additive model. The formula of computing τ_h is given by

$$\tau_h = 1 - \frac{\rho_{YXh}^2 + \rho_{YZh}^2 - 2\rho_{YXh}\rho_{YZh}\rho_{XZh}}{1 - \rho_{XZh}^2}$$

where ρ_{YXh} and ρ_{XZh} are respectively the correlation coefficients between Y and X ; X and Z in the h^{th} stratum.

Table 5: Privacy protection measure τ_h

Stratum No.	$\tau_{h(Linear.Model)}$	$\tau_{h(M_1)}$	$\tau_{h(M_2)}$
1	0.1745063	0.1762518	0.001828369
2	0.1424378	0.1431637	0.001794860
3	0.2012377	0.2017166	0.001431154
4	0.2118853	0.2123574	0.001672627
5	0.1955643	0.1958186	0.001583843
6	0.1677808	0.1678604	0.001523790

Remark 2: It is mentioned that if the value of τ_h becomes higher, more privacy is protected. Furthermore, $\tau_h = 1$ means maximum privacy protection and $\tau_h = 0$ means no privacy is protected.

Tables 2, 3 and 4 reveal that the suggested calibration estimator with scrambled response on second call under non-response provides much better results than that of the Hansen and Hurwitz (1946) estimator. From the tables 2, 3 and 4, we can also develop a relationship as given below:

$$MSE(\hat{\bar{y}}_{HH(cal)M_1}) > MSE(\hat{\bar{y}}_{HH(cal)}) > MSE(\hat{\bar{y}}_{HH(cal)M_2})$$

From the above relationship, we can conclude that the proposed calibration estimator is highly efficient under the scrambled model M_2 given by Pollock and Bek (1976). But, it is well known fact that the estimator which is high in efficiency may be less in privacy protection. So, in terms of privacy protection, the proposed calibration estimator under scrambled model M_1 can be a better choice. Table 5 shows that τ_h under the scrambled model M_1 is more close to 1 which indicates that the estimator under the scrambled model M_1 is rich in privacy protection whereas τ_h under the scrambled model M_2 is more close to 0 which indicates that less privacy is protected under the scrambled model M_2 .

6. CONCLUDING REMARKS

We have proposed a calibration estimator for estimating the mean of a stratified population under non-response using scrambled response on second call for those who are selected in the sub-sample. The information on a single auxiliary variable has been utilized to develop the calibration estimator. To scramble the response, we have considered the scrambled linear model given by Diana et al. (2014). We have used the MSE as a tool to measure the efficiency of the proposed calibration estimator. To strengthen the results, we have also computed the privacy protection. To measure the amount of privacy protection, we have used the normalized privacy protection measure given by Diana et al. (2013). To see the trade-off between efficiency and privacy protection, we have used two more scrambled models given by Saha (2007) and Pollock and Bek (1976). We have compared the efficiency of the proposed calibration estimator with that of Hansen and Hurwitz (1946) estimator under scrambled response on second call. From the tables 2, 3, 4 and 5, it is concluded that we get more précised estimates by using Pollock and Bek (1976) additive model and if we use Saha (2007) mixed model for scrambling the response, more privacy is protected.

7. REFERENCES

- Ahmed, S., Shabbir, J. & Hafeez, W., 2016, Hansen and Hurwitz estimator with scrambled response on second call in stratified random sampling. *Journal of Statistics Applications & Probability* **5(2)**, 357-369.
- Chaudhary, M. K., Singh, R., Kumar, M., Singh, A. K., & Adewara, A. A., 2012, A class of estimators for assessing the population mean under non-response using auxiliary information. *ASM Sc. J* **6(1)**, 23-30.
- Diana, G., & Perri, P. F., 2009, Estimating a sensitive proportion through randomized response procedures based on auxiliary information. *Statistical Papers* **50(3)**, 661-672.
- Diana, G., & Perri, P. F., 2010, New scrambled response models for estimating the mean of a sensitive quantitative character. *Journal of Applied Statistics* **37(11)**, 1875-1890.
- Diana, G., Giordan, M., & Perri, P. F., 2013, Randomized response surveys: A note on some privacy protection measures. *Model Assisted Statistics and Applications* **8(1)**, 19-28.
- Diana, G., Riaz, S., & Shabbir, J., 2014, Hansen and Hurwitz estimator with scrambled response on the second call. *Journal of Applied Statistics* **41(3)**, 596-611.
- Eriksson, S. A., 1973, A new model for randomized response. *International Statistical Review/Revue Internationale de Statistique*, 101-113.
- Fox, J. A., & Tracy, P. E., 1986, *Randomized response: A method for sensitive surveys* (No. 58). SAGE Publications, Incorporated.
- Greenberg, B. G., Abul-El, A. L. A., Simmons, W. R., & Horvitz, D. G., 1969, The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association* **64(326)**, 520-539.

Hansen, M. H., & Hurwitz, W. N., 1946, The problem of non-response in sample surveys. *Journal of the American Statistical Association* **41(236)**, 517-529.

Khare, B. B. & Pandey, S. K., 2000, A class of estimators for ratio of two population means using auxiliary character in presence of non-response. *J. Sc. Res.* **50**, 115-124.

Mangat, N. S., & Singh, R., 1990, An alternative randomized response procedure. *Biometrika* **77(2)**, 439-442.

Pollock, K. H., & Bek, Y., 1976, A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association* **71(356)**, 884-886.

Saha, A., 2007, A simple randomized response technique in complex surveys. *Metron* **65(1)**, 59-66.

Shabbir, J., & Gupta, S., 2005, On modified randomized device of Warner's model. *Pakistan Journal of Statistics-All Series-* **21(1)**, 123.

Shabbir, J., & Khan, N. S., 2013, On estimating the finite population mean using two auxiliary variables in two phase sampling in the presence of non response. *Communications in Statistics-Theory and Methods* **42(22)**, 4127-4145.

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