PLOS ONE

RESEARCH ARTICLE

Optimum second call imputation in PPS sampling

Fariha Sohil[1], **Muhammad Umair Sohail**[2]*, **Javid Shabbir**[3]

1 Department of Education, The Women University, Multan, Pakistan, **2** Department of Statistics, University of Narowal, Narowal, Pakistan, **3** Department of Statistics, Quaid-I-Azam University, Islamabad, Pakistan

* umairsohailch@gmail.com

OPEN ACCESS

Citation: Sohil F, Sohail MU, Shabbir J (2022) Optimum second call imputation in PPS sampling. PLoS ONE 17(1): e0261834. https://doi.org/10.1371/journal.pone.0261834

Editor: Dejan Dragan, Univerza v Mariboru, SLOVENIA

Received: June 19, 2021

Accepted: December 12, 2021

Published: January 21, 2022

Data Availability Statement: In this research, a hypothetical data set is used which can be easily regenerated at the given value of parameters with the help of available statistical software. The parameters are included in the paper and its Supporting information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The current study deals with imputation of item non-response in probability proportional to size (PPS) sampling. A new imputation procedure is proposed by using the known co-variance between the study variable and the auxiliary variable in the case of quantitative sensitive study variable by considering the non-response in a randomization mechanism on the second call. An empirical study is conducted at the optimum values of k_{og} and n_{og} for the relative comparisons of ratio, difference, and proposed estimators, respectively, with the Hansen-Hurwitz estimator.

1 Introduction

Survey sampling is a technique which is utilizes in almost every field of life to estimate the finite population parameters with limited response. There are many sample selection procedures, which provide reliable data by selecting the representative sample. In equal probability sampling schemes, the probability of selection is equal for all the units in target population. If units varying in size, equal probability sampling may not give the appropriate importance to large or small units in the population. The appropriate importance to the population units is assigned by allocating the unequal probabilities of selection to the different units in the population. Thus, when units are different in size and variable under study is correlated with their auxiliary information e.g. size, then the selection probabilities may be assigned in proportion to their sizes. For example,

1. Colleges with large number of educational departments are likely to have more students and more faculty members. For the funds allocation, it may well be desirable to adopt a scheme of selection in which colleges are selected with probabilities proportional to their students or departments.
2. In an industrial survey, the number of workers may be selected as size of industrial area.
3. In biological studies, the number of patients may be selected according to the size of the hospital.

For all of these cases, the selection of sampling units is proportional to the size of auxiliary information associated with the particular unit, is called sampling with probability proportional to size (*PPS*). It is well known that the proper use of auxiliary information at estimation

stage or at design stage or at both stages is helpful to magnify the performance of resultant estimators. Ratio, product, and regression estimators are good examples in this context.

In many real life situations, where non-response/refusals may affect the reliability and accuracy of data sets. These refusals are mostly occurred due to many reasons such as time of survey (during summer or winter vacations, office hours etc.), survey contents (embarrassing nature of questions, double barrel question etc.), respondent burden (irrelevant questions, length of questionnaire etc.), or data collection methods (telephone or mail surveys, personal interviews etc).

Initially, [1] provides an idea of sub-sampling the non-respondents of first call by dividing the population into two strata; respondents and non-respondents at first call. The detailed discussion on the proposed estimator is given in Subsections 1.1 and 1.2 for the case of simple random and PPS sampling scheme, respectively.

1.1 Sample selection in simple random sampling

Let $\Omega = \{\Omega_1, \Omega_2, \Omega_3, \cdots, \Omega_N\}$ be a finite population of N units. Let y_i and (x_i, z_i) be the values of the study variable (y) and the auxiliary variable (x, z), respectively, for $i = 1, 2, \cdots, N$. Assume that x_i has high positive and z_i has low positive correlation, respectively, with the study variable (y_i). So, x_i is used at the estimation stage and z_i is used at the sample selection stage from population. Let a sample $\{(\Im = \Im_1, \Im_2, \cdots, \Im_n)\}$ of size n be selected using simple random sampling without replacement(SRSWOR) scheme. Assume that n_{1s} units respond at first call, report their responses $y_{i(1)}$ and n_{2s} units do not respond at first call. Further, a sample of size $r_{1s} = \frac{n_{2s}}{k}$, where $k > 1$, is drawn from n_{2s} non-respondent group, report their responses $y_{i(2)}$, belong to group G_1 and $r_{2s} = r_{1s}(k-1)$ are those, who refuse to report their response belong to group G_2.

Thus, the sub-sampling estimate for population mean, is given by

$$\begin{aligned} \bar{y}^* &= \frac{1}{n}\left\{\sum_{i=1}^{n_{1s}} y_{i(1)} + \sum_{i=1}^{n_{2s}} y_{i(2)}\right\} \\ &= w_{1s}\bar{y}_{(1)} + w_{2s}\bar{y}_{(2)}, \end{aligned} \tag{1}$$

where $w_{1s} = \frac{n_{1s}}{n}$, $w_{2s} = \frac{n_{2s}}{n}$, $\bar{y}_{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$, $\bar{y}_{(1)} = \frac{1}{r_1}\sum_{i=1}^{n_2} y_i$ and r be the respondents. The variance of $\bar{y}^*$ is given by

$$V(\bar{y}^*) = \frac{(N-n)}{nN}\bar{Y}^2 C_y^2 + \frac{1}{n}W_2(k-1)\bar{Y}_2^2 C_{y_{(2)}}^2, \tag{2}$$

where $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$, $C_y^2 = \frac{S_y^2}{\bar{Y}^2}$, $\bar{Y}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} y_i$, $S_{y(2)}^2 = \frac{1}{N_2-1}\sum_{i=1}^{N_2}(y_i - \bar{Y}_2)^2$, $C_{y(2)}^2 = \frac{S_{y(2)}^2}{\bar{Y}_2^2}$, and $W_2 = \frac{N_2}{N}$.

1.2 Selection of sample with PPS sampling

In PPS sampling scheme, the selection of units in the sample is carried with probability proportional to a given measure of size, where the size is measured by the available suitable auxiliary information. Let $u_i = y_i/(N\pi_i)$ and $v_i = x_i/(N\pi_i)$, where $\pi_i = z_i \Big/ \sum_{i=1}^{N} z_i$ and also let $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$ and $\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$ be the unbiased estimators of population means and their variances are $V(\bar{u}) = \sum_{i=1}^{n} p_i(u_i - \bar{u})^2$ and $V(\bar{v}) = \sum_{i=1}^{n} p_i(v_i - \bar{v})^2$, respectively, where $p_i = z_i \Big/ \sum_{i=1}^{n} z_i$. It is also assumed that the average value of u_i is approximately equal to average value of y_i.

Let a sample $\{s_i = (s_1, s_2, s_3, \cdots, s_n)\}$ of size n be selected using PPS with replacement sampling scheme. Assume that n_1 units respond at first call, report their responses $u_{i(1)} = y_{i(1)}/(N_1 \pi_{1i})$, where $\pi_{1i} = z_i \Big/ \sum_{i=1}^{N_1} z_i$ and n_2 units do not respond at first call. Further, a sample of size, $r_1 = \frac{n_2}{k}$, is drawn from n_2 non-respondent group, report their responses $u_{i(2)} = y_{i(2)}/(N_2\pi_{2i})$, where $\pi_{2i} = z_i \Big/ \sum_{i=1}^{N_2} z_i$, belongs to group G_1 and $r_2 = r_1(k-1)$ is those, who refuse to report their responses belong to group G_2. Thus, the Hansen-Hurwitz estimator under PPS sampling scheme can be modified as:

$$\begin{aligned} \bar{u}^* &= \frac{1}{n}\left\{\sum_{i=1}^{n_1} u_{i(1)} + \sum_{i=1}^{n_2} u_{i(2)}\right\} \\ &= \frac{1}{n}\{n_1\bar{u}_{(1)} + n_2\bar{u}_{(2)}\} \\ &= w_1\bar{u}_{(1)} + w_2\bar{u}_{(2)}, \end{aligned} \tag{3}$$

where n_1 and r_1 are the PPS respondent units at first and second calls, respectively. The variance of $\bar{u}^*$ is given by

$$V(\bar{u}^*) = \frac{(N-n)}{nN}\bar{Y}^2 C_u^2 + \frac{1}{n} W_2(k-1)\bar{Y}_2^2 C_{u_{(2)}}^2, \tag{4}$$

where $S_u^2 = \sum_{i=1}^{N} \pi_i (u_i - \bar{Y})^2$, $S_{u_{(2)}}^2 = \sum_{i=1}^{N_2} \pi_{2i}(u_{i(2)} - \bar{Y}_2)^2$, $C_u^2 = \frac{S_u^2}{\bar{Y}^2}$, and $C_{u_{(2)}}^2 = \frac{S_{u(2)}^2}{\bar{Y}_2^2}$.

1.3 Statement of the problem

When variables of interest are sensitive or embarrassing in nature, then respondents are reluctant to report their true responses or may refuse to respond. Several statistical models are available in literature to protect the confidentiality and privacy of interviewee by hiding their identities, which are helpful to reduce the non-response bias. A pioneer idea of randomized response technique (RRT) was described by [2] to handle the high rate of refusals due to sensitive nature of questions. Commonly, these refusals have been occurred during the analysis of demographic and economic variables, respectively, etc. Interest readers may be referred to read [3–9], and many others. [10, 11] use the randomized response models (RRMs) for obtaining the true status of interviewee on second attempt. The proposed estimators by these researchers can perform better as compared to traditional ones.

The aim of this investigation is to study the missing complete at random (MCAR) values at second call, when the interviewees are reluctant to use RRMs. For the non-respondents of first call, different additive, multiplicative and subtractive models, respectively, might be utilized to create the feeling among respondents that their privacy is secured beside their truthful response.

For creating privacy protection felling among non-respondents of first call, we consider to modify linear randomized response model proposed by [12]. From the n_2 non-respondents of first call, the scrambled response is obtained using the [12] model.

1.3.1 Privacy protection at second call. Let the i^{th} respondent draw two cards i.e S_{1i} and S_{2i} from two independent decks of cards, say D_1 and D_2, respectively, which are un-correlated with y. At the second call, the i^{th} respondent can report the scrambled response as follows:

$$u'_{i(2)} = S_{1i}u_{i(2)} + S_{2i} \tag{5}$$

Let E_3 and V_3 be, respectively, the expected value and variance over the scrambled device. We assume that $E_3(S_{1i}) = \theta_1$, $E_3(S_{2i}) = \theta_2$, $V_3(S_{1i}) = \sigma_1^2$ and $V_3(S_{2i}) = \sigma_2^2$ with $E_3(u'_{i(2)}) =$

$\theta_1 u_{i(2)} + \theta_2$ and $V_3(u'_{i(2)}) = \sigma_1^2 u_{i(2)}^2 + \sigma_2^2$. Also let $\hat{u}_{i(2)}$ be the suitable transformation of randomized response for the i^{th} unit whose expectation under (5) model coincides with the true response y_i, as:

$$\hat{u}_{i(2)} = \frac{\theta_2}{\theta_1}\left(\frac{u'_{i(2)}}{\theta_2} - 1\right) \tag{6}$$

with

$$V_3\left(\hat{u}_{i(2)}\right) = C_{\theta_1}^2 \hat{u}_i^2 + \left(\frac{\theta_2}{\theta_1}\right)^2 C_{\theta_2}^2, \tag{7}$$

where $C_{\theta_1}^2 = \frac{\sigma_1^2}{\theta_1^2}$ and $C_{\theta_2}^2 = \frac{\sigma_2^2}{\theta_2^2}$.

At the second call, out of n_2 non-respondent of first call, only r_1 interviewees can give their scrambling responses and remaining r_2 units cannot give their true or scrambled responses. Let $\hat{\bar{u}}_{(r_1)} = \frac{1}{r_1}\sum_{i=1}^{r_1} \hat{u}_{i(r_1)}$ be the sample mean of respondent class at second attempt.

2 Modifying existing literature

In this section, we modify the exiting literature as per the statement of the problem. The most commonly used imputation procedures are discussed in Subsection 2.1, 2.2, and 2.3.

2.1 Mean estimator

In this section, our focus is to impute the missing r_2 values by using conventional method of imputation. The missing structure is defined as follows:

$$\Delta_i = \begin{cases} \hat{u}_{i(r_1)} & \text{if the } i^{th} \text{ respondent report their scrambled response} \\ \hat{\bar{u}}_{(r_1)} & \text{otherwise} \end{cases} \tag{8}$$

Hence, the whole population is divided in $\Omega_{(1)}$ and $\Omega_{(2)}$ strata having N_1 and N_2 units, respectively. Furthermore, $\Omega_{(2)}$ is divided into two groups G_1 and G_2 of size R_1 and R_2 units, respectively, when N_1, N_2, R_1 and R_2 are known in advance. For the case of scrambled responses at second call, the point Hansen-Hurwitz estimator for population mean $(\bar{Y})$ can be modified as:

$$\begin{aligned} \hat{\bar{u}}^* &= \frac{1}{n}\sum_{i=1}^{n}\Delta_i \\ &= \frac{1}{n}\left\{\sum_{i=1}^{n_1}\Delta_i + \sum_{i=1}^{n_2}\Delta_i\right\} \\ &= \frac{1}{n}\left\{n_1\bar{u}_{(1)} + \sum_{i=1}^{r_1}u_{i(r_1)} + \sum_{i=1}^{r_2}\hat{\bar{u}}_{(r_1)}\right\} \\ &= \frac{1}{n}\left\{n_1\bar{u}_1 + r_1\hat{\bar{u}}_{(r_1)} + (n_2 - r_1)\hat{\bar{u}}_{(r_1)}\right\} \\ &= w_1\bar{u}_{(1)} + w_2\hat{\bar{u}}_{(r_1)} \end{aligned} \tag{9}$$

So, we have the following Lemmas.

Lemma 2.1 The variance of $\hat{\bar{u}}_{(r_1)}$, is given by

$$V\{\hat{\bar{u}}_{(r_1)}\} = \frac{1}{N_2 r_1}\sum_{i=1}^{N_2}\phi_i + \left(\frac{1}{r_1} - \frac{1}{N_2}\right)\bar{Y}_2^2 C_{u_{(2)}}^2. \tag{10}$$

Proof. **Proof**: Let E_j and V_j, $j = (1, 2)$ be the expected values and variances for given n_2 and r_1, respectively. Then, by the definition of variance, we have

$$\begin{aligned}
V\{\hat{\bar{u}}_{(r_1)}\} &= E_1E_2V_3\{\hat{\bar{u}}_{(r_1)}\} + E_1V_2E_3\{\hat{\bar{u}}_{(r_1)}\} + V_1E_2E_3\{\hat{\bar{u}}_{(r_1)}\} \\
&= E_1E_2V_3\left\{\frac{1}{r_1}\sum_{i=1}^{r_1}\hat{u}_i\right\} + E_1V_2E_3\left\{\frac{1}{r_1}\sum_{i=1}^{r_1}\hat{u}_i\right\} + V_1E_2E_3\left\{\frac{1}{r_1}\sum_{i=1}^{r_1}\hat{u}_i\right\} \\
&= E_1E_2\left\{\frac{1}{r_1^2}\sum_{i=1}^{r}V_3(\hat{u}_i)\right\} + E_1V_2\left\{\frac{1}{r_1}\sum_{i=1}^{r_1}E_3(\hat{u}_i)\right\} + V_1E_2\left\{\frac{1}{r_1}\sum_{i=1}^{r}E_3(\hat{u}_i)\right\} \\
&= E_1E_2\left[\frac{1}{r_1^2}\sum_{i=1}^{r_1}\left\{C_{\theta_1}^2\hat{u}_i^2 + \left(\frac{\theta_2}{\theta_1}\right)^2 C_{\theta_2}^2\right\}\right] + E_1V_2\left\{\frac{1}{r_1}\sum_{i=1}^{r}u_{i(2)}\right\} \\
&\quad + V_1E_2\left\{\frac{1}{r_1}\sum_{i=1}^{r_1}u_{i(2)}\right\} \\
&= E_1\left[\frac{1}{n_2r_1}\sum_{i=1}^{n_2}\left\{C_{\theta_1}^2\left(\frac{N_2-1}{N_2}S_{u_{(2)}}^2 + \bar{Y}_2^2\right) + \left(\frac{\theta_2}{\theta_1}\right)^2 C_{\theta_2}^2\right\}\right] \\
&\quad + E_1\left(\frac{1}{r_1} - \frac{1}{n_2}\right)s_{u_{(2)}}^2 + V_1\left\{\frac{1}{n_2}\sum_{i=1}^{n_2}u_{(2)}\right\} \\
&= \frac{1}{N_2r_1}\sum_{i=1}^{N_2}\{\bar{Y}_2^2C_{\theta_1}^2\left(1 + \frac{N_2-1}{N_2}C_{u_{(2)}}^2\right) + \left(\frac{\theta_2}{\theta_1}\right)^2 C_{\theta_2}^2\} + \left(\frac{1}{r_1} - \frac{1}{n_2}\right)\bar{Y}_2^2C_{u_{(2)}}^2 \\
&\quad + \left(\frac{1}{n_2} - \frac{1}{N_2}\right)\bar{Y}_2^2C_{u_{(2)}}^2. \\
&= \frac{1}{N_2r_1}\sum_{i=1}^{N_2}\{\bar{Y}_2^2C_{\theta_1}^2\left(1 + \frac{N_2-1}{N_2}C_{u_{(2)}}^2\right) + \left(\frac{\theta_2}{\theta_1}\right)^2 C_{\theta_2}^2\} + \left(\frac{1}{r_1} - \frac{1}{N_2}\right)\bar{Y}_2^2C_{u_{(2)}}^2 \\
&= \frac{1}{N_2r_1}\sum_{i=1}^{N_2}\phi_i + \left(\frac{1}{r_1} - \frac{1}{N_2}\right)\bar{Y}_2^2C_{u_{(2)}}^2.
\end{aligned} \tag{11}$$

Corollary 2.1.1. It is important to note that $V\{\hat{\bar{u}}_{(r_1)}\}$ requires the second moment (μ_{2u}) of y, which is generally unknown. [13] suggested two possible ways to acquire μ_{2u}: (i) guess it from the prior information or pilot survey and (ii) obtain the sample estimate to derive the information about μ_{2u} by keeping in mind the sensitive nature of u_i.

Lemma 2.2. The variance of $\hat{\bar{u}}^*$, is given by

$$V(\hat{\bar{u}}^*) = \frac{k}{Nnr_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}W_2(k-1)\bar{Y}_2^2C_{u_{(2)}}^2 + \frac{(N-n)}{nN}\bar{Y}^2C_u^2. \tag{12}$$

Proof. **Proof** Let E_m and V_m, $m = (4, 5)$ be the expected values and variances for given N_1 and N_2, respectively. By definition, we have

$$\begin{aligned} V(\hat{\bar{u}}^*) &= E_4V_5\{\hat{\bar{u}}^*|n_1,n_2\} + V_4E_5\{\hat{\bar{u}}^*|n_1,n_2\} \\ &= E_4V_5\left\{\frac{w_1\bar{u}_{(1)} + w_2\hat{\bar{u}}_{(r_1)}}{n}\middle|n_1,n_2\right\} + V_4E_5\left\{\frac{w_1\bar{u}_{(1)} + w_2\hat{\bar{u}}_{(r_1)}}{n}\middle|n_1,n_2\right\} \\ &= E_4\left\{\frac{1}{N_2r_1}\sum_{i=1}^{N_2}\phi_i + \frac{n_2^2}{n^2}\left(\frac{1}{r_1}-\frac{1}{n_2}\right)s^2_{y_{(2)}}\right\} + V_4\{\bar{u}_{(n)}\} \\ &= E_4\{\frac{1}{N_2r_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}\left(\frac{n_2}{r_1}-1\right)\frac{n_2}{n}s^2_{u_{(2)}}\} + V_4\{\bar{u}_{(n)}\} \\ &= \frac{k}{Nnr_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}W_2(k-1)\bar{Y}_2^2C^2_{u_{(2)}} + \frac{(N-n)}{nN}\bar{Y}^2C_u^2. \end{aligned} \tag{13}$$

By ignoring correction factor $\left(1-\frac{n}{N}\right)$ for the ease of computation, then we have

$$V\left(\hat{\bar{u}}^*\right) = \frac{k}{Nnr_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}W_2(k-1)\bar{Y}_2^2C^2_{u_{(2)}} + \frac{1}{n}\bar{Y}^2C_u^2. \tag{14}$$

Corollary 2.2.1. From (4) and (14), we see that the variance of modified estimator is higher than Hansen-Hurwitz estimator. It means that $\hat{\bar{u}}^*$ is less efficient than $\bar{u}^*$.

The objective of our study is to increase the truth and confidence among interviewees that their privacy is secure beside their true answers. Moreover, the non-response at first call might be occurred due to non-availability or inability to provide the required information. Therefore, at the second call, it may happen that those people are willing to report their responses directly, even the sensitive characteristics are investigated. For this purpose, the randomization in stages should be re-expounded as an optional randomized response (ORR) procedure, which permits the respondents to divulging the direct or true response without using RRT, is given by

$$\hat{u}_i = t_iu_i + (1-t_i)u_{i(1)}, \tag{15}$$

where

$$t_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ respondent report their direct response} \\ 0 & \text{otherwise} \end{cases}$$

It is easy to show that the unbiased estimator for $\bar{Y}$ is derived by replacing (15) in (9) and its variance becomes $(1 - t_i)\phi_i$ instead of ϕ_i, in (14). Furthermore, ORR reduces the variance and privacy at various values of t_i for the non-respondents at first call.

2.2 Ratio estimator

Initially, [14] takes into account the utility of auxiliary information at estimation stage by defining the ratio estimator for population. The traditional ratio estimator can be modified for the imputation of missing scrambled responses at second call, as:

$$\triangle_{i(c)} = \begin{cases} \hat{u}_{i(r_1)} & \text{if} \quad i \in G_1 \\ \dfrac{1}{1-f_1}\left(\hat{\bar{u}}_{(r_1)}\dfrac{\bar{X}_2}{\bar{v}_{(r_1)}} - f_1\hat{\bar{u}}_{(r_1)}\right) & \text{if} \quad i \in G_2 \end{cases}, \tag{16}$$

where $f_1 = \frac{r_1}{n_2}$, $\bar{v}_{r_1} = \frac{1}{r_1}\sum_{i=1}^{r_1}\frac{x_i}{p_{2i}}$, and $\bar{X}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2}x_i$.

The point estimator for sub-population ($\Omega_{(2)}$), is given by

$$\begin{aligned}\hat{\bar{u}}_{(c)} &= \frac{1}{n_2}\sum_{i=1}^{n_2}\Delta_{i(c)} \\ &= \frac{1}{n_2}\left\{\sum_{i=1}^{r_1}\Delta_{i(c)} + \sum_{i=1}^{r_2}\Delta_{i(c)}\right\} \\ &= \frac{1}{n_2}\left\{\sum_{i=1}^{r_1} y_{i(r_1)} + \frac{1}{1-f_1}\sum_{i=1}^{r_2}\left(\hat{\bar{u}}_{(r_1)}\frac{\bar{X}_2}{\bar{v}_{r_1}} - f_1\hat{\bar{u}}_{(r_1)}\right)\right\} \\ &= \frac{1}{n_2}\{r_1\hat{\bar{u}}_{(r_1)} + n_2\hat{\bar{u}}_{(r_1)}\frac{\bar{X}_2}{\bar{v}_{(r_1)}} - r_1\hat{\bar{u}}_{(r_1)}\} \\ &= \hat{\bar{u}}_{(r_1)}\frac{\bar{X}_2}{\bar{v}_{(r_1)}}\end{aligned} \tag{17}$$

The Hansen-Hurwitz ratio estimator for population mean ($\bar{Y}$), is given by

$$\hat{\bar{u}}^*_{(c)} = w_1\bar{u}_{(1)} + w_2\hat{\bar{u}}_{(c)} \tag{18}$$

The variance of modified ratio estimator is given by

$$\begin{aligned}V(\hat{\bar{u}}^*_{(c)}) &= E_4V_5\{\hat{\bar{u}}^*_{(c)}|n_1,n_2\} + V_4E_5\{\hat{\bar{u}}^*_{(c)}|n_1,n_2\} \\ &= E_4V_5\left\{\frac{w_1\bar{u}_{(1)} + w_2\hat{\bar{u}}_{(c)}}{n}\middle|n_1,n_2\right\} + V_4E_5\left\{\frac{w_1\bar{u}_{(1)} + w_2\hat{\bar{u}}_{(c)}}{n}\middle|n_1,n_2\right\} \\ &= E_4\left\{\frac{1}{N_2r_1}\sum_{i=1}^{N_2}\phi_i + \frac{n_2^2}{n^2}\left(\frac{1}{r_1}-\frac{1}{n_2}\right)s^2_{r_{(2)}}\right\} + V_4\{\bar{u}_{(n)}\} \\ &= E_4\left\{\frac{1}{N_2r_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}\left(\frac{n_2}{r_1}-1\right)\frac{n_2}{n}s^2_{r_{(2)}}\right\} + V_4\{\bar{u}_{(n)}\} \\ &= \frac{k}{Nnr_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}W_2(k-1)\bar{Y}_2^2C^2_{r_{(2)}} + \frac{1}{n}\bar{Y}_2^2C_y^2,\end{aligned} \tag{19}$$

where $C^2_{r_{(2)}} = C^2_{u_{(2)}} + C^2_{v_{(2)}} - 2\rho_{uv}C_{u_{(2)}}C_{v_{(2)}}$,

$$C^2_{u_{(2)}} = \frac{S^2_{u_{(2)}}}{\bar{Y}_2},\ C^2_{v_{(2)}} = \frac{S^2_{v_{(2)}}}{\bar{X}_2}, \rho_{uv} = \frac{S_{uv_{(2)}}}{S_{u_{(2)}}S_{v_{(2)}}}, S^2_{u_{(2)}} = \frac{1}{N_2-1}\sum_{i=1}^{N_2}\pi_{2i}(u_i-\bar{Y}_2)^2,$$

$$S^2_{v_{(2)}} = \frac{1}{N_2-1}\sum_{i=1}^{N_2}\pi_{2i}(v_i-\bar{X}_2)^2, S_{uv_{(2)}} = \frac{1}{N_2-1}\sum_{i=1}^{N_2}\pi_{2i}(u_i-\bar{Y}_2)(v_i-\bar{X}_2).$$

2.3 Difference estimator

Now, we consider the difference estimator for explaining missing structure of scrambled responses, as:

$$\triangle_{i(d)} = \begin{cases} \hat{u}_{i(r_1)} & \text{if} \quad i \in G_1 \\ \dfrac{1}{1-f_1}\{\hat{\bar{u}}_{(r_1)} + d\left(\bar{X}_2 - \bar{v}_{(r_1)}\right) - f_1\hat{\bar{u}}_{(r_1)}\} & \text{if} \quad i \in G_2 \end{cases}, \tag{20}$$

where d is an unknown constant.

The point estimator for sub-population mean (Ω), is given by

$$\begin{aligned} \hat{\bar{u}}_{(d)} &= \frac{1}{n_2}\sum_{i=1}^{n_2}\Delta_{i(d)} \\ &= \frac{1}{n_2}\left\{\sum_{i=1}^{r_1}\Delta_{i(d)} + \sum_{i=1}^{r_2}\Delta_{i(d)}\right\} \\ &= \frac{1}{n_2}\left\{\sum_{i=1}^{r_1}u_{i(r_1)} + \frac{1}{1-f_1}\sum_{i=1}^{r_2}\left(\hat{\bar{u}}_{(r_1)} + d\left(\bar{X}_2 - \bar{v}_{(r_1)}\right) - f_1\hat{\bar{u}}_{(r_1)}\right)\right\} \\ &= \frac{1}{n_2}\{r_1\hat{\bar{u}}_{(r_1)} + n_2\{\hat{\bar{u}}_{(r_1)} + d\left(\bar{X}_2 - \bar{v}_{(r_1)}\right)\} - r_1\hat{\bar{u}}_{(r_1)}\} \\ &= \hat{\bar{u}}_{(r_1)} + d(\bar{X}_2 - \bar{v}_{(r_1)}) \end{aligned} \tag{21}$$

The combined version of modified Hansen-Hurwitz estimator is given by

$$\hat{\bar{u}}^*_{(d)} = w_1\bar{u}_{(1)} + w_2\hat{\bar{u}}_{(d)} \tag{22}$$

The variance of $\hat{\bar{u}}^*_{(d)}$ estimator, is stated as

$$\begin{aligned} V(\hat{\bar{u}}^*_{(d)}) &= E_4V_5\{\hat{\bar{u}}^*_{(d)}|n_1,n_2\} + V_4E_5\{\hat{\bar{u}}^*_{(d)}|n_1,n_2\} \\ &= E_4V_5\left\{\frac{w_1\bar{u}_{(1)} + w_2\hat{\bar{u}}_{(d)}}{n}\middle| n_1,n_2\right\} + V_4E_5\left\{\frac{w_1\bar{u}_{(1)} + w_2\hat{\bar{u}}_{(d)}}{n}\middle| n_1,n_2\right\} \\ &= E_4\left\{\frac{1}{N_2r_1}\sum_{i=1}^{N_2}\phi_i + \frac{n_2^2}{n^2}\left(\frac{1}{r_1} - \frac{1}{n_2}\right)s^2_{d_{(2)}}\right\} + V_4\{\bar{u}_{(n)}\} \\ &= E_4\left\{\frac{1}{N_2r_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}\left(\frac{n_2}{r_1} - 1\right)\frac{n_2}{n}s^2_{d_{(2)}}\right\} + V_4\{\bar{u}_{(n)}\} \\ &= \frac{k}{Nnr_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}W_2(k-1)C^2_{d_{(2)}} + \frac{1}{n}\bar{Y}^2_2C^2_y, \end{aligned} \tag{23}$$

where $C^2_{d_{(2)}} = \bar{Y}^2_2C^2_{u_{(2)}} + d^2\bar{X}^2_2C^2_{v_{(2)}} - 2k\bar{Y}_2\bar{X}_2\rho_{uv_{(2)}}C_{u_{(2)}}C_{v_{(2)}}$.

When $d_{\text{opt.}} = \frac{\bar{Y}_2}{\bar{X}_2}\frac{C_{u_{(2)}}}{C_{v_{(2)}}}\rho_{uv_{(2)}}$, variance of $\hat{\bar{u}}^*_{(d)}$ reduces to

$$V\left(\hat{\bar{u}}^*_{(d)}\right) \cong \frac{1}{Nnr_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}W_2(k-1)\bar{Y}^2_2C^{2*}_{d_{(2)}} + \frac{1}{n}\bar{Y}^2C^2_u, \tag{24}$$

where $C^{2*}_{d_{(2)}} = C^2_u(1-\rho^2_{uv_{(2)}})$.

The problem of estimating the population parameters by using higher order moments of the auxiliary variable was considered by [15–17]. Later on [18–20] among others, also contemplate the known higher order moments of the auxiliary variable for estimation of finite population parameters. In the theory of survey sampling, it is well established result that the use of higher order moments of the auxiliary variable plays a pivotal role in estimating the finite population mean of the study variable. This literature inspired the researchers to impute the missing values at second call by using known covariance between the study variable and the auxiliary variable.

3 Proposed imputation procedure

Initially, [21] improves the conventional mean estimator by using a tuning constant ($\alpha_{(s)}$), in the case of missing values, as:

$$\triangle_{i(s)} = \begin{cases} \alpha_{(s)}\hat{u}_{i(r_1)} & \text{if} \quad i \in G_1 \\ \alpha_{(s)}\hat{\bar{u}}_{(r_1)} & \text{if} \quad i \in G_2 \end{cases}, \tag{25}$$

which leads to Searls's type estimator for $\bar{u}_{(2)}$ is given by

$$\bar{y}_{(s)} = \alpha_{(s)}\hat{\bar{u}}_{(r_1)}. \tag{26}$$

Although Searls's approach uses the known coefficient of variation to increase the efficiency of the estimation procedure. The optimum value of $\alpha_{(s)}$ depends on $C_{u_{(2)}}$, $C_{v_{(2)}}$ and $\rho_{uv_{(2)}}$, which are stable quantities. The stability of these constant has been explored by numerous researchers like [22–24], etc. Therefor, the present investigation is a significant search of optimum imputation method by using the co-variance between the study and auxiliary variable. The imputation of item non-response is given by

$$\triangle_{i(p)} = \begin{cases} \alpha_1\hat{u}_{i(r_1)} & \text{if} \quad i \in G_1 \\ \alpha_1\hat{\bar{u}}_{(r_1)} + \dfrac{\alpha_2}{1-f_1}\left(\bar{X}_2 - \bar{v}_{(r_1)}\right) + \dfrac{\alpha_3}{1-f_1}\left(S_{uv_{(2)}} - s_{uv_{(r_1)}}\right) & \text{if} \quad i \in G_2 \end{cases}, \tag{27}$$

where α_1, α_2, and α_3 are suitable chosen constants and are determined by minimizing the resultant mean square error. The point estimator for population mean, is defined as:

$$\begin{aligned} \hat{\bar{u}}_{(p)} &= \frac{1}{n_2}\sum_{i=1}^{n_2}\Delta_{i(p)} \\ &= \frac{1}{n_2}\left\{\sum_{i=1}^{r_1}\Delta_{i(p)} + \sum_{i=1}^{r_2}\Delta_{i(p)}\right\} \\ &= \frac{1}{n_2}\left\{\sum_{i=1}^{r_1}\alpha_1 u_{i(r_1)} + \sum_{i=1}^{r_2}\left(\alpha_1\hat{\bar{u}}_{(r_1)} + \frac{\alpha_2}{1-f_1}\left(\bar{X}_2 - \bar{v}_{(r_1)}\right) + \frac{\alpha_3}{1-f_1}\left(S_{uv_{(2)}} - s_{uv_{(r_1)}}\right)\right)\right\} \\ &= \frac{1}{n_2}\left\{\alpha_1 r_1\hat{\bar{u}}_{(r_1)} + r_2\left(\alpha_1\hat{\bar{u}}_{(r_1)} + \frac{\alpha_2}{1-f_1}\left(\bar{X}_2 - \bar{v}_{(r_1)}\right) + \frac{\alpha_3}{1-f_1}\left(S_{uv_{(2)}} - s_{uv_{(r_1)}}\right)\right)\right\} \\ &= \alpha_1\hat{\bar{u}}_{(r_1)} + \alpha_2(\bar{X}_2 - \bar{v}_{(r_1)}) + \alpha_3(S_{uv_{(2)}} - s_{uv_{(r_1)}}) \end{aligned} \tag{28}$$

The modified version of Hansen-Hurwitz difference estimator is given by

$$\hat{\bar{u}}^{*}_{(d)} = w_1 \bar{u}_{(1)} + w_2 \hat{\bar{u}}_{(p)} \tag{29}$$

The variance of $\hat{\bar{u}}_{(p)}$, is given by

$$\begin{aligned} V(\hat{\bar{u}}^{*}_{(p)}) &= E_4 V_5\left\{\hat{\bar{u}}^{*}_{(p)}|n_1, n_2\right\} + V_4 E_5\left\{\hat{\bar{u}}^{*}_{(p)}|n_1, n_2\right\} \\ &= E_4 V_5\left\{\frac{w_1 \bar{u}_{(1)} + w_2 \hat{\bar{u}}_{(p)}}{n}|n_1, n_2\right\} + V_4 E_5\left\{\frac{w_1 \bar{u}_{(1)} + w_2 \hat{\bar{u}}_{(p)}}{n}|n_1, n_2\right\} \\ &= E_4\left\{\frac{1}{N_2 r_1}\sum_{i=1}^{N_2}\phi_i + \frac{n_2^2}{n^2}\left(\frac{1}{r_1} - \frac{1}{n_2}\right)s^2_{p_{(2)}}\right\} + V_4\left\{\bar{u}_{(n)}\right\} \\ &= E_4\left\{\frac{1}{N_2 r_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}\left(\frac{n_2}{r_1} - 1\right)\frac{n_2}{n}s^2_{p_{(2)}}\right\} + V_4\left\{\bar{u}_{(n)}\right\} \\ &= \frac{k}{Nnr_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}W_2(k-1)S^2_{p_{(2)}} + \frac{1}{n}\bar{Y}_2^2 C_y^2, \end{aligned} \tag{30}$$

where

$$\begin{aligned} S^2_{p_{(2)}} &= \alpha_1^2 \bar{Y}_2^2 C^2_{u_{(2)}} + \alpha_2^2 \bar{X}^2 C^{2*}_{v_{(2)}} + \alpha_3^2 S^2_{uv_{(2)}}\left(\frac{\lambda_{22_{(2)}}}{\rho_{uv_{(2)}}} - 1\right) - 2\,\alpha_1\alpha_2 \bar{Y}_2 \bar{X}_2 \rho_{uv_{(2)}} C_{u_{(2)}} C_{v_{(2)}} \\ &\quad -2\,\alpha_1\alpha_3 \bar{Y}_2 S_{uv_{(2)}} C_{u_{(2)}}\frac{\lambda_{21_{(2)}}}{\rho_{uv}} + 2\,\alpha_2\,\alpha_3\,\bar{Y}_2 S_{uv_{(2)}} C_{u_{(2)}}\frac{\lambda_{21_{(2)}}}{\rho_{uv_{(2)}}} + \bar{Y}_2^2(\alpha_1 - 1)^2, \\ \lambda_{ab_{(2)}} &= \frac{\mu_{ab_{(2)}}}{\mu^{a/2}_{20_{(2)}}\mu^{b/2}_{02_{(2)}}},\ \mu_{ab_{(2)}} = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2}\pi_{2i}(u_{i_{(2)}} - \bar{Y}_2)^a(v_{i_{(2)}} - \bar{X}_2)^b. \end{aligned}$$

The optimum values of α_j, $j = (1, 2, 3)$ are obtained, respectively, by minimizing (30), as follows:

$$\begin{aligned} \alpha_{1(\text{opt.})} &= \frac{1}{1 + \chi^2_{(2)}}, \\ \alpha_{2(\text{opt.})} &= \frac{\bar{Y}_2 C_{u_{(2)}}\left\{\lambda_{21_{(2)}}\lambda_{12_{(2)}} - \varphi_{(2)}\rho_{uv_{(2)}}\right\}\alpha_{1(\text{opt.})}}{\bar{X}_2 C_{v_{(2)}}\left\{-\varphi_{(2)} + \lambda^2_{12_{(2)}}\right\}}, \quad \text{and} \\ \alpha_{3(\text{opt.})} &= \frac{\bar{Y}_2 C^*_{u_{(2)}}\rho^2_{uv}\left\{\lambda_{12_{(2)}} - \frac{\lambda_{21_{(2)}}}{\rho^2_{uv_{(2)}}}\right\}\alpha_{1(\text{opt.})}}{S_{uv_{(2)}}\left\{-\varphi_{(2)} + \lambda^2_{12}\right\}}, \end{aligned} \tag{31}$$

where $\chi^2_{(2)} = C^{2*}_{u_{(2)}}(1 - Q^2_{u.vs_{uv_{(2)}}})$, $\varphi_{(2)} = \rho^2_{uv_{(2)}}\left(\frac{\lambda_{22_{(2)}}}{\rho^2_{uv_{(2)}}} - 1\right)$, and $Q^2_{u.vs_{uv_{(2)}}} = \left\{\lambda^2_{21_{(2)}} + \rho^2_{uv_{(2)}}\varphi_{(2)} - 2\rho^*_{uv_{(2)}}\lambda_{21_{(2)}}\lambda_{12_{(2)}}\right\}(\varphi_{(2)} - \lambda^2_{12_{(2)}})^{-1}$ is the coefficient of multiple determination of u on v and $s_{uv_{(2)}}$.

Substituting (31) in (30), the variance of $\hat{\bar{u}}_{(p)}$, is given by

$$V(\hat{\bar{u}}^{*}_{(p)})_{\min.} \cong \frac{1}{Nnr_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}W_2(k-1)\bar{Y}_2^2 C^{2*}_{p_{(2)}} + \frac{1}{n}\bar{Y}C_u^2, \tag{32}$$

where $C^{2*}_{p_{(2)}} = \chi^2_{(2)}\left\{1+\chi^2_{(2)}\right\}^{-1}$.

Remark 1. The second term in $V(\hat{\bar{u}}^{*}_{(p)})_{\min.}$ is vanished, if $k = 1$. It happens when each non-respondent of first call is interviewed at second call.

4 Choice of sampling fractions

We shall deduce the optimum values of k and n that minimize the variance at specified cost. The cost function for the proposed model is based on following four components, as:

1. C_0 = over head cost.
2. C_1 = per unit cost for collecting the response by mail inquiry at first call.
3. C_2 = the unit cost for obtaining the scrambled response from the non-respondent group of first call.
4. C_3 = cost per unit for editing, processing or imputing the missing r_2 values.

Thus, the cost function is given by

$$C^{*} = nC_0 + n_1C_1 + r_1C_2 + r_2C_3 \tag{33}$$

Note that C^{*} is the total cost, thus it varies from sample to sample. So, we use the expected cost by applying the expectation on (33), we have

$$\begin{aligned} E(C^{*}) &= E\{nC_0 + n_1C_1 + r_1C_2 + r_2C_3\} \\ &= n\left\{C_0 + E\left(\frac{n_1}{n}\right)C_1 + E\left(\frac{r_1}{n}\right)C_2 + E\left(\frac{r_2}{n}\right)C_3\right\} \\ &= \frac{n}{N}\left\{NC_0 + N_1C_1 + \frac{1}{k}(R_1C_2 + R_2C_3)\right\} \\ &= n\left\{C_0 + W_1C_1 + \frac{1}{Nk}C_{(2)}\right\} \end{aligned} \tag{34}$$

So, we have the following Lemma, as:

Lemma 4.1. The optimum values of k and n for the minimum expected cost are, respectively, given by

$$k_{ov} = \sqrt{\frac{C_{(2)}\left(\bar{Y}^2C_u^2 + \frac{1}{Nr_1}\sum_{i=1}^{N_2}\phi_i - W_2\bar{Y}_2^2C^{2*}_{g_{(2)}}\right)}{N(C_0 + W_1C_1)W_2\bar{Y}_2^2C^{2*}_{g_{(2)}}}} \tag{35}$$

and

$$n_{ov} = \frac{\bar{Y}^2 C_u^2 + \frac{1}{Nr_1}\sum_{i=1}^{N_2}\phi_i + W_2(k-1)\bar{Y}_2^2 C_{g_{(2)}}^{2*}}{V_0 + \frac{\bar{Y}^2 C_y^2}{N}}, \tag{36}$$

where $g = c$, d, and p.

Proof. Let the variance $V(\hat{\bar{u}}_{(g)})$ be a fixed V_0, i.e $V(\hat{\bar{u}}_{(g)}) = V_0$, then the Lagrange function, is given by

$$L \quad = n\left\{C_0 + W_1 C_1 + \frac{1}{Nk}C_{(2)}\right\} + \xi\left\{V(\hat{\bar{u}}_{(g)})_{\min.} - V_0\right\}, \tag{37}$$

where ξ is a Lagrange multiplier. Differentiating (37) with respect to n, equating to zero i.e $\left(\frac{\partial L}{\partial n} = 0\right)$ and ignoring δ_j, j = (1or2). We have

$$\frac{\partial L}{\partial n} = C_0 + W_1 C_1 + \frac{1}{Nk}C_{(2)} + \frac{\xi}{n^2}\left\{-\bar{Y}^2 C_u^2 - \frac{1}{Nr_1}\sum_{i=1}^{N_2}\phi_i - W_2(k-1)\bar{Y}_2^2 C_{g_{(2)}}^{2*}\right\} = 0$$

which implies

$$n = \frac{\sqrt{\xi}\sqrt{N\left\{\bar{Y}^2 C_u^2 + \frac{1}{Nr_1}\sum_{i=1}^{N_2}\phi_i + W_2(k-1)\bar{Y}_2^2 C_{g_{(2)}}^{2*}\right\}}}{\sqrt{N\left(C_0 + W_1 C_1 + \frac{1}{Nk}C_{(2)}\right)}} \tag{38}$$

Note that

$$V(\hat{\bar{u}}^*_{(g)})_{\min.} \cong \frac{1}{Nnr_1}\sum_{i=1}^{N_2}\phi_i + \frac{1}{n}W_2(k-1)\bar{Y}_2^2 C_{g_{(2)}}^{2*} + \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}^2 C_u^2 \tag{39}$$

Substituting (38) in (39), we have

$$\sqrt{\xi} = \frac{\sqrt{N\left(C_0 + W_1 C_1 + \frac{1}{Nk}C_{(2)}\right)}\sqrt{\bar{Y}^2 {C_u}^2 + \frac{1}{Nr_1}\sum_{i=1}^{N_2}\phi_i + W_2(k-1)\bar{Y}_2^2 C_{g_{(2)}}^{2*}}}{\sqrt{N}\left(V_0 + \frac{1}{N}\bar{Y}^2 C_u^2\right)} \tag{40}$$

Substituting (40) in (38), we have

$$n_{ov} = \frac{\bar{Y}^2 C_y^2 + \frac{1}{Nr_1}\sum_{i=1}^{N_2}\phi_i + W_2(k-1)\bar{Y}_2^2 C_{g_{(2)}}^{2*}}{V_0 + \frac{\bar{Y}^2 C_u^2}{N}} \tag{41}$$

which is the required optimum sample size (n_{ov}). Now, we differentiate (37) with respect to k and equate to zero i.e $\left(\frac{\partial L}{\partial k} = 0\right)$. Then, we have

$$\frac{\partial L}{\partial k} = \xi\frac{1}{n}W_2\bar{Y}_2^2 C_{g_{(2)}}^{2*} - \frac{1}{Nk^2}C_{(2)} = 0$$

which implies

$$k^2 = \frac{nC_{(2)}}{N\xi\, W_2 \bar{Y}_2^2 C_{g_{(2)}}^2} \tag{42}$$

Using (38) in (42), we have

$$k_{ov} = \sqrt{\frac{C_{(2)}\left(\bar{Y}^2 C_u^2 + \frac{1}{Nr_1}\sum_{i=1}^{N_2}\phi_i - W_2\bar{Y}_2^2 C_{g_{(2)}}^{2*}\right)}{N(C_0 + W_1 C_1) W_2 \bar{Y}_2^2 C_{g_{(2)}}^{2*}}} \tag{43}$$

which is the required optimum value of k.

Corollary 4.1.1. The optimum values of n and k are proportional to the expected cost (C^*). To get optimum values of k and n, that, $V(\hat{\bar{u}}_{(g)}^*)_{\min.}$, we simply substitute C_y^2 and $C_{g_{(2)}}^2$ in (35) and (36).

5 Empirical comparison

On the lines of [25], the relative comparison of $\hat{\bar{u}}_{(q)}^*$ with respect to $\hat{\bar{u}}^*$ is considered by generating a hypothetical population under following key steps, as:

1. Let two independent populations say $\{\Omega_{(x)}$ and $\Omega_{(z)}\}$ of size 1000 are obtained from gamma distribution, using following parametric values, as:

$$\gamma_x = \begin{pmatrix} 1.5 \\ 5.5 \end{pmatrix} \quad \text{and} \quad \gamma_z = \begin{pmatrix} 1.5 \\ 4.5 \end{pmatrix} \tag{44}$$

The study variable is generated by

$$y_i = 2.8 + \sqrt{(1-0.2)}z_i + x_i, \quad \rho_{uv} = 0.85 \tag{45}$$

2. Splitting the populations into two strata having $N_1 = 690$ and $N_2 = 310$ units.
3. Assume that, out of N_2 units, R_1 units provides the response by using (5) and remaining $R_2 = (N_2 - R_1)$ are those who refuse to give their true or scrambled responses.
4. Imputing the missing R_2 values by using $\bar{X}$ and S_{uv}.
5. Repeat the process $H = \binom{N}{n_{ov}}$ times. The variance of the given estimator is obtained by using following expression, as:

$$V\left(\hat{\bar{u}}_{(g)}^*\right) = \frac{1}{H}\sum_{l=1}^{H}\left\{\hat{\bar{u}}_{(v)}^*(l) - \bar{Y}\right\}^2 \tag{46}$$

and the relative efficiency (R.E) of $\hat{\bar{u}}_{(g)}^*$ is obtained by using the following expression

$$R.E(j) = \frac{V(\hat{\bar{u}}^*)}{V(\hat{\bar{u}}_{(g)}^*)} \quad \text{for} \quad j = 1, 2, \text{ and } 3. \tag{47}$$

For the numerical comparison, we consider the following values of un-known constants, as:

$$C^{*}=200,\ C_{o}=20,\ C_{1}=50,\ C_{\theta_1}=C_{\theta_2}=10,\ \left(\frac{\theta_2}{\theta_1}\right)^{2}=0.1,\ \text{and}\ r=0.4\,n_{og},$$

where r is assumed response rate, which is 40% of n_{og}. The optimum values of relative efficiencies (R.Es) of $\hat{\bar{u}}^{*}_{(g)}$ are given in Table 1.

Table 1 shows the optimum values of k, n, and R.E(j) of estimators i.e modified ratio and difference estimators. Under this hypothetical population, the modified estimators i.e $\hat{\bar{u}}^{*}_{(g)}$ perform better as compared to traditional Hansen-Hurwitz estimator $(\hat{\bar{u}}^{*})$. We also observe that the optimum value of n_{og} is approximately similar for all $\hat{\bar{u}}^{*}_{(p)}$, so optimum sample of size n_{op} is used for the relative comparison between existing and proposed imputation estimators.

From Table 1, we observed following proportionality relationships between C_2, C_3, r_1, V_o, k_{op}, n_{op}, and R.E(j).

1. The values of V_o and n_{op} have inverse relationship with C_2 and C_3.
2. C_2 and C_3 have the positive relationship with RE(j). As the costs of scrambling response and imputation increase, the relative efficiencies of $\hat{\bar{u}}^{*}_{(g)}$ have been improved significantly.

Table 1. Optimum values of k, n, and R.E(j) w.r.t. $\hat{\bar{u}}^{*}$.

C_2	C_3	r_1	V_o	k_{op}	n_{op}	Relative Efficiency		
						R.E(1)	R.E(2)	R.E(3)
1	2.0	10	4	2.30	191	1.48	2.08	3.83
1	2.0	20	8	1.63	58	1.75	2.72	3.88
1	2.0	30	12	1.34	28	1.91	3.41	3.96
1	1.0	10	4	1.93	191	1.34	2.08	3.37
1	1.0	20	8	1.37	58	1.93	3.01	3.64
1	1.0	30	12	1.12	28	1.64	2.42	3.91
1	0.5	10	4	1.71	191	1.54	1.67	3.68
1	0.5	20	8	1.21	58	1.56	2.88	4.01
1	0.5	30	12	1.09	28	1.22	1.95	4.52
2	2.0	10	4	2.68	191	1.67	2.40	3.89
2	2.0	20	8	1.90	58	1.88	3.39	4.45
2	2.0	30	12	1.56	28	2.49	3.57	4.89
2	1.0	10	4	2.38	191	1.57	1.85	3.29
2	1.0	20	8	1.69	58	1.26	2.91	3.90
2	1.0	30	12	1.38	28	1.57	2.85	4.04
2	0.5	10	4	2.29	191	1.59	2.23	3.06
2	0.5	20	8	1.57	58	1.47	2.74	3.19
2	0.5	30	12	1.29	28	1.33	2.65	3.45
3	2.0	10	4	2.99	191	1.36	1.86	2.44
3	2.0	20	8	2.12	58	1.59	2.28	3.22
3	2.0	30	12	1.74	28	1.85	2.61	4.02
3	1.0	10	4	2.74	191	1.48	2.59	2.74
3	1.0	20	8	1.94	58	1.66	2.97	3.07
3	1.0	30	12	1.59	28	1.84	3.72	4.36
3	0.5	10	4	2.61	191	1.61	2.49	3.36
3	0.5	20	8	1.85	58	1.78	3.16	3.54
3	0.5	30	12	1.51	28	2.33	4.11	4.31

https://doi.org/10.1371/journal.pone.0261834.t001

3. r_1 has the negative association with k_{op} and n_{op}. The values of k_{op} and n_{op} decrease as r_1 increasing.
4. V_o also has the inverse relationship with k_{op} and n_{op}. As the value of V_o increases, the values of k_{op} and n_{op} decrease.
5. The relative efficiencies of $\hat{\bar{u}}^*_{(g)}$ are also inversely correlated with r_1 and V_o. The values of R. E(j) decrease as V_o and n_{op} increase.

From the numerical finding, we can conclude that the proposed imputation procedure at second call should be performs better as compared to existing and tradition Hansen-Hurwitz estimators at various values of C_2, C_3, r_1 and V_o.

6 Conclusion

The problem of non-response bias in the sensitive quantitative study variable has been diminished by sub-sampling the non-respondent, viz. Hansen and Hurwitz (1946) procedure. A new imputation mechanism has been defined by using the known co-variance between the study variable and the auxiliary variable. Optimum value for sample size is also derived for a given set of unit cost (C_q, $q = 0, 1, 2, 3$), r_1, and V_o. From the Table 1, we can easily say that the proposed imputation method can outperforms as compared to ratio, difference, and Hansen-Hurwitz estimators.

When the processing, editing, or imputing cost per unit is high, the proposed imputation strategy can performs better as compared to their counterpart. Our proposed imputation procedure is also useful when there are serious concerns about the non-response bias or refusals due to the sensitive nature of the study variable that is difficult to ignore it.

Supporting information

S1 Code. In this research a hypothetical data set is used which can be easily regenerated at the given value of parameters with the help of available statistical software.
(R)

Acknowledgments

We are grateful to the reviewers and the associate editor for their in depth comments for improving the quality of the article.

Author Contributions

Conceptualization: Fariha Sohil.

Data curation: Muhammad Umair Sohail.

Formal analysis: Muhammad Umair Sohail.

Funding acquisition: Muhammad Umair Sohail.

Investigation: Javid Shabbir.

Methodology: Fariha Sohil.

Project administration: Muhammad Umair Sohail.

Resources: Muhammad Umair Sohail.

Software: Muhammad Umair Sohail.

Supervision: Muhammad Umair Sohail.

Validation: Fariha Sohil.

Visualization: Fariha Sohil.

Writing – original draft: Fariha Sohil.

Writing – review & editing: Muhammad Umair Sohail, Javid Shabbir.

References

1. Hansen M. H. and Hurwitz W. N. (1946). The problem of non-response in sample surveys. Journal of the American Statistical Association, 41(236):517–529 https://doi.org/10.1080/01621459.1946.10501894 PMID: 20279350
2. Abul-Ela A.-L. A., Greenberg G. G., and Horvitz D. G. (1967). A multi-proportions randomized response model. Journal of the American Statistical Association, 62(319):990–1008. https://doi.org/10.1080/01621459.1967.10500910
3. Warner S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309):63–69. https://doi.org/10.1080/01621459.1965.10480775 PMID: 12261830
4. Greenberg B. G., Abul-Ela A.-L. A., Simmons W. R., and Horvitz D. G. (1969). The unrelated question randomized response model: Theoretical framework. Journal of the American Statistical Association, 64(326):520–539. https://doi.org/10.1080/01621459.1969.10500991
5. Moors J. (1971). Optimization of the unrelated question randomized response model. Journal of the American Statistical Association, 66(335):627–629. https://doi.org/10.1080/01621459.1971.10482320
6. Folsom R. E., Greenberg B. G., Horvitz D. G., and Abernathy J. R. (1973). The two alternate questions randomized response model for human surveys. Journal of the American Statistical Association, 68(343):525–530. https://doi.org/10.1080/01621459.1973.10481377
7. Eichhorn B. H. and Hayre L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. Journal of Statistical Planning and Inference, 7(4):307–316. https://doi.org/10.1016/0378-3758(83)90002-2
8. Mangat N. and Singh R. (1990). An alternative randomized response procedure. Biometrika, pages 439–442. https://doi.org/10.1093/biomet/77.2.439
9. Gjestvang C. R. and Singh S. (2006). A new randomized response model. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(3):523–530. https://doi.org/10.1111/j.1467-9868.2006.00554.x
10. Diana G., Riaz S., and Shabbir J. (2014). Hansen and hurwitz estimator with scrambled response on the second call. Journal of Applied Statistics, 41(3):596–611. https://doi.org/10.1080/02664763.2013.846305
11. Ahmed S., Shabbir J., and Gupta S. (2017). Use of scrambled response model in estimating the finite population mean in presence of non response when coefficient of variation is known. Communications in Statistics-Theory and Methods, 46(17):8435–8449. https://doi.org/10.1080/03610926.2016.1183782
12. Diana G. and Perri P. F. (2010a). New scrambled response models for estimating the mean of a sensitive quantitative character. Journal of Applied Statistics, 37(11):1875–1890. https://doi.org/10.1080/02664760903186031
13. Diana G. and Perri P. F. (2010b). New scrambled response models for estimating the mean of a sensitive quantitative character. Journal of Applied Statistics, 37(11):1875–1890. https://doi.org/10.1080/02664760903186031
14. Cochran W. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. The Journal of Agricultural Science, 30(02):262–275. https://doi.org/10.1017/S0021859600048012
15. Srivastava S. K. and Jhajj H. S. (1981). A class of estimators of the population mean in survey sampling using auxiliary information. Biometrika, 68(1):341–343. https://doi.org/10.1093/biomet/68.1.341
16. Isaki C. T. (1983). Variance estimation using auxiliary information. Journal of the American Statistical Association, 78(381):117–123. https://doi.org/10.1080/01621459.1983.10477939
17. Singh S. and Horn S. (1998). An alternative estimator for multi-character surveys. Metrika, 48(2):99–107.

18. Mohamed C., Sedory S. A., and Singh S. (2016). Imputation using higher order moments of an auxiliary variable. Communications in Statistics-Simulation and Computation, 46(8):6588–6617. https://doi.org/10.1080/03610918.2016.1208235

19. Sohail M. U., Shabbir J., and Ahmed S. (2017). Modified class of ratio and regression type estimators for imputing scrambling response. Pakistan Journal of Statistics, 33(4):277–300.

20. Bhushan S., Pratap Pandey A., and Pandey A. (2018). On optimality of imputation methods for estimation of population mean using higher order moment of an auxiliary variable. Communications in Statistics-Simulation and Computation, pages 1–15.

21. Searls D. T. (1964). The utilization of a known coeffcient of variation in the estimation procedure. Journal of the American Statistical Association, 59(308):1225–1226. https://doi.org/10.1080/01621459.1964.10480765

22. Murthy M. N. (1967). Sampling theory and methods. Calcutta-35: Statistical Publishing Society, 204/1, Barrackpore Trunk Road, India.

23. Reddy V. (1978). A study on the use of prior knowledge on certain population parameters in estimation. Sankhya C, 40:29–37.

24. Singh S. (2009). A new method of imputation in survey sampling. Statistics, 43(5):499–511. https://doi.org/10.1080/02331880802605114

25. Okafor F. C. and Hyunshik L. (2000). Double sampling for ratio and regression estimation with sub-sampling the non-respondents. Survey Methodology, 26(2):183–188.