Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/120115

Estimation of Mean in Two-Staged Unequally Clustered Population in Presence of Non-Response Situation

P. Mukhopadhyay[1,*] *and A. Bandyopadhyay*[1] *and G.N. Singh* [2]

[1]Department of Mathematics, Asansol Engineering College, Asansol-713305, India
[2]Department of Mathematics & Computing, IIT (ISM) Dhanbad, Dhanbad-826004, India

Received: 21 Nov. 2021, Revised: 21 Mar. 2022, Accepted: 23 May 2022
Published online: 1 Jan. 2023

Abstract: In this article, authors have tried to implement a modified version of [1] estimator in a two-stage unequally clustered population where the second stage units can be either of responding or non-responding type. We have utilized imputation technique to tackle the non-response persisting even after second call in the situation of two-staged unequal cluster population which is completely a new attempt. The purpose is to strengthen the new estimator by improving the 'second call' part of the estimator. Instead of taking a simple weighted average of the responding part after 'second call' a difference type imputation (to deal with the non-responding part yet after 'second call' in a different way) is deployed to extrapolate the average and thus, examine the performance of the proposed strategy. Empirical studies carried over the data set of natural population. Suitable recommendations to the survey statistician are made.

Keywords: Two-stage Cluster Sampling, non-response, difference type imputation, sub sampling of the non-responding group

1 Introduction

Cluster sampling is a popular sampling technique where the entire population is divided into clusters and a random sample of the selected clusters are first included in the sample as because sometimes it is not possible to draw a sample of ultimate units of interest on the first occasion since such units is not readily accessible. However, a list of some suitably defined bigger units or first stage units (FSU's) is prepared from each of which a sample (SSU's) may be selected. Instead of completely enumerating all the SSU's belonging to the selected FSU's as in case of cluster sampling, one may select sample of SSU's from the list of all SSU's belonging to the selected FSU's . The sampling is therefore carried out in two stages and therefore, it is called two stage sampling. If a non-sensitive question is asked there might be still some cases of non-response due to system failure. A phone call may fail, the person may be out of station, e-mail sent may bounce for some reason etc. Hansen and Hurwitz first came out with a model for a flat population where in the second attempt some non-respondents changed to respondents and the simple weighted average of the mean value of study variable obtained in this phase is added to the overall estimator. The [1] estimator for a population while non-response exists after both attempts takes the form:

$$\bar{Y}_{HH} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}_2' = \frac{1}{m}\sum_{j=1}^{m} y_j \tag{1}$$

where n_2 units remains non-responding after first attempt and $m = \frac{n_2}{k}(k > 1)$ is the integral faction who did respond after second attempt.

As a real life example of random non-response we may think of the situation as follows:

Suppose a professional survey team is trying to assess the support base of a political party X in a target area (may be a state or area under municipal corporation etc.). It is always convenient to list some zone wise residences address first (FSUs) and then interact with the households leaving at those addresses (SSUs). A typical voter may not disclose his/her

* Corresponding author e-mail: parthamukhopadhyay1967@gmail.com

choice at first call over mail/questionnaire methods but at personal interaction (second call) he/she may speak out. Some individual may never respond (due to fear or other reasons). Even some may change their choices during the process of survey due to some very unexpected events took place in between influencing their thought process. In this paper, we have taken an attempt to improve the second part in two ways. Firstly, instead of accepting the simple weighted average we go through an imputation process to project the mean response of the finally non-responding part from the knowledge of mean value of the responding part at second call. Moreover we have implemented it to a two stage clustered population of unequal SSU sizes. Almost no work is done in this direction to the best of author's knowledge. A difference type imputation is chosen and an auxiliary variable x is considered whose mean value is known at all level. The procedures and formulations for this are discussed in the sections.

The non-response of many respondents renders sample returned incomplete. To deal with missing values systematically, [2,3] suggested imputation methods that make incomplete data sets structurally complete. Imputation can also be carried out with the aid of an auxiliary variable. For example, [4] used the information on an available auxiliary variable for imputation purpose. Further, [5] and [6] suggested several imputation based methods utilizing the available auxiliary information to deal with the problems of non-response in sample surveys.

The core of the present work lies in utilizing the imputation method to extract the actual proportion of sensitive character from the non-respondents even after second call in the situation of two-staged unequal cluster population. (in practice there are hardly any real life population where number of SSU units are same in all FSU 's e.g., the households as FSU 's and the number of members (the SSU 's) in each household which may vary randomly. Such populations are often studied for various socio-economic surveys). The new estimator is shown to be much superior after improving the 'second call' part of the estimator. A simple weighted average (usually taken) of the responding part after 'second call' is replaced by a robust difference type imputation. The conclusion in this paper can be summarized by saying that imputation is giving always a better output than the conventional [1] version in two-stage clustered population with unequal SSU sizes.

2 Description of Population and Formulation of Sample Structure:

We shall implement here sub sampling of non-responding units adopted earlier by [1] in a two-stage unequal population. For that we take a finite population U be partitioned into N first stage units (FSU) denoted by $(U_1, U_2, \cdots, U_N)$. The number of second stage units (SSU) in ith FSU be M_i. (where M_{i1} is the assumed population responding part and M_{i2} the non-responding in first call so that $M_i = M_{i1} + M_{i2}$). Let be the values of the study and auxiliary variables respectively on the jth SSU ($j = 1,2,...,M_i$) within the ith($i = 1,2,\cdots,N$).Further, we also assume that the information on x is available at all levels.

Further, we consider the occurrence of non-response situation in the following ways.

To estimate the overall population mean $\left(\bar{Y}_{..} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{M_i}\sum_{i=1}^{M_i} Y_{ij}\right)\right)$ of the study variable y, a sample $S_1 (S_1 \subset U)$ of size n FSU 's is drawn out of N FSU 's by SRSWOR. Then a second stage sample (SSU) S_2 of size $m_i (< M_i)$ is drawn from the ith FSU and approached to answer the sensitive question(first call) in a quantitative way. Let m_{i1} responded and m_{i2} did not ($m_i = m_{i1} + m_{i2}, mi1 \leq M_{i1}$ and $m_{i2} \leq M_{i2}$). Next a second approach is arranged (may be personal interaction or anything alike) for the non-responding part m_{i2} and a subsample of integral size $m_{i2r} = m_{i2}/k (k > 1, m_{i2r} < m_{i2} < m_i)$ are found to response this time. The procedure is followed for each of n FSUs sampled.

It is to be noted that the sample for second call per FSU is drawn as a subsample of that drawn on first call. We mention here that the [1] version of estimator for the ith FSU while non-response exists even after second call takes the form:

$$\bar{Y}_{iHH} = \frac{m_{i1}}{m_i}\bar{y}_{i.1} + \frac{m_{i2}}{m_i}\bar{y}'_{i.2} \text{ where } \bar{y}'_{i.2} = \frac{1}{m_{i2r}}\sum_{j=1}^{m_{i2r}} y_{ij} \tag{2}$$

2.1 Notations Used to Formulate the Estimator and Its Variance

Hence onwards, we use the following notations for population and sample parameters:

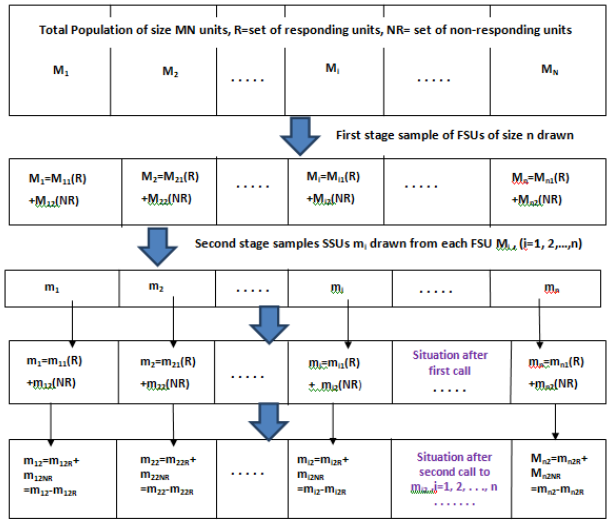

Fig. 1: Pictorial Representation of the Sampling Procedure

2.1.1 For the population

$$
\begin{aligned}
Y_{..} &= \sum_{i=1}^{N}\left(\sum_{j=1}^{N} Y_{ij}\right) : \text{ over all population total of the study variable y} \\
X_{..} &= \sum_{i=1}^{N}\left(\sum_{j=1}^{N} X_{ij}\right) : \text{ over all population total of the study variable x} \\
M_0 &= \sum_{i=1}^{N} M_i : \text{ total number of ssu units in population}
\end{aligned} \tag{3}
$$

$$
M_1 = \sum_{i=1}^{N} M_{i1}, M_2 = \sum_{i=1}^{N} M_{i2} : \text{total responding SSUs units (assumed) in population (1st, 2nd call)}
$$

$$
\text{for equal SSU size } Y_{..} = \sum_{i=1}^{N}\left(\sum_{j=1}^{N} Y_{ij}\right), X_{..} = \sum_{i=1}^{N}\left(\sum_{j=1}^{N} X_{ij}\right)
$$

$$
\begin{aligned}
\bar{Y}_{..} &= \frac{1}{N}\sum_{1}^{N}\left(\frac{1}{M_i}\frac{1}{M_i}\sum_{1}^{M_i} Y_{ij}\right) = \frac{1}{N}\sum_{1}^{N}\bar{Y}_i : \text{ Overall population mean of y} \\
\bar{X}_{..} &= \frac{1}{N}\sum_{1}^{N}\left(\frac{1}{M_i}\frac{1}{M_i}\sum_{1}^{M_i} X_{ij}\right) = \frac{1}{N}\sum_{1}^{N}\bar{X}_i : \text{ Overall population mean of y}
\end{aligned} \tag{4}
$$

$$\bar{Y}_{i.} = \frac{1}{M_i}\sum_{j=1}^{M_i} Y_{ij} : \text{ population mean (y) for ith FSU}$$

$$\bar{X}_{i.} = \frac{1}{M_i}\sum_{j=1}^{M_i} X_{ij} : \text{ population mean (x) for ith FSU}$$

$$S'^2_{yb} = \frac{1}{N-1}\sum i^N(\bar{Y}_{i.} - \bar{Y}_{..})^2 : \text{ population variances (y) between FSU-means}$$

$$S'^2_{xb} = \frac{1}{N-1}\sum i^N(\bar{X}_{i.} - \bar{X}_{..})^2 : \text{ population variances (x) between FSU-means}$$

(we shall replace S'^2_{yb} by S^2_{yb} for equal SSU size case)

$$S'^2_{ywi} = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(Y_{ij} - \bar{Y}_{i.})^2 : \text{ population variances (y) within ith FSU}$$

$$S'^2_{xwi} = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(X_{ij} - \bar{X}_{i.})^2 : \text{ population variances (x) within ith FSU} \tag{5}$$

$\bar{S}'^2_{yw} = \frac{1}{N}\sum_1^N S^2_{ywi}, \bar{S}'^2_{xw} = \frac{1}{N}\sum_1^N S^2_{xwi}$: mean population variances (y,x) within ith FSU

$$\rho^2_{yxwi} = \{\frac{1}{M_i - 1}\sum_{j=1}^{M_i}(Y_{ij} - \bar{Y}_{i.})(X_{ij} - \bar{X}_{i.})\}/\sqrt{S'^2_{xwi}S'^2_{ywi}}$$

$$\rho^2_{yxb} = \{\frac{1}{N-1}\sum_{j=1}^{N}(Y_{i.} - \bar{Y}_{..})(X_{i.} - \bar{X}_{..})\}/\sqrt{S^2_{xB}S^2_{yB}}$$

$$\rho^2_{yx2wi} = \{\frac{1}{M_{i2} - 1}\sum_{j=1}^{M_{i2}}(Y_{ij} - \bar{Y}_{2i.})(X_{ij} - \bar{X}_{2i.})\}/\sqrt{S'^2_{xw2i}S'^2_{yw2i}}$$

$$S'^2_{yw2i} = \frac{1}{M_{i2} - 1}\sum_{j=1}^{M_{i2}}(Y_{ij} - \bar{Y}_{2i.})^2$$

$$f_1 = \frac{n}{N}, f_{2i} = \frac{m_i}{M_i}, \bar{Y} = \frac{Y}{N}, M_0 = \sum_{i=1}^{N} M_i = NM \text{ (for equal SSU size)} \tag{6}$$

$W_2 =$ avg. non-response ratio ($\frac{M_2}{M_0}$) after first call in population- kept same across homogeneous clusters.

Value of W_2 may be found from previous experience. Here taken 0.5 to 0.35 typically.

2.1.2 For sample:

$$
\begin{aligned}
y_{i.} &= \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij} : \text{sample mean(y) on ith FSU} \\
x_{i.} &= \frac{1}{m_i}\sum_{j=1}^{m_i} x_{ij} : \text{sample mean(x) on ith FSU} \\
&\text{based on the respondent region in } S_1(R). \\
\bar{y}_{i.1} &= \frac{1}{m_{i1}}\sum_{j=1}^{m_{i1}} y_{ij1} : \text{ sample mean (y) of ith FSU after first call} \\
\bar{y}_{i.2r} &= \frac{1}{m_{i2r}}\sum_{j=1}^{m_{i2r}} y_{ij2r} : \text{ sample mean (y) of ith FSU based on the respondents after 2nd call in} S_2(R^c) \\
\bar{y}_{i(mi2-mi2/k)} &= \frac{1}{m_{i2}-m_{i2r}}\sum_{j=1}^{m_i-m_{i2r}} y_{ij} \\
s'^2_{ywi} &= \frac{1}{m_i-1}\sum_{1}^{m_i}(y_{ij}-\bar{y}_{i.})^2 : \text{ sample variance (y) within FSU} \\
s'^2_{xwi} &= \frac{1}{m_i-1}\sum_{1}^{m_i}(x_{ij}-\bar{x}_{i.})^2 : \text{ sample variance (x) within FSU} \\
\bar{s}'^2_{yw} &= \frac{1}{n}\sum_{1}^{n} s'^2_{ywi}, \bar{s}'^2_{xw} = \frac{1}{n}\sum_{1}^{n} s'^2_{xwi} \\
s'^2_{yb} &= \frac{1}{n-1}\sum_{1}^{n}(\bar{y}_{i.}-\bar{y}_{..}))^2, s'^2_{xb} = \frac{1}{n-1}\sum_{1}^{n}(\bar{x}_{i.}-\bar{x}_{..}))^2 : \text{ sample variance between FSU- means} \\
&\text{(we replaces}'^2_{yb} \text{ by } s^2_{yb} \text{ for equal SSU size case)} \\
\rho^2_{yxwi} &= \{\frac{1}{m_i-1}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i.})(x_{ij}-\bar{x}_{i.})\}/s'^2_{xwi} \text{ correlation coefficients} \\
\rho^2_{yxb} &= \{\frac{1}{n-1}\sum_{j=1}^{n}(y_{i.}-\bar{y}_{..})(x_{i.}-\bar{x}_{..})\}/\sqrt{s'^2_{xb}s'^2_{yb}} \\
\rho^2_{yx2b} &= \{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_{2i.}-\bar{y}_{..})(\bar{x}_{2i.}-\bar{x}_{..})\}/\sqrt{s'^2_{x2b}s'^2_{y2b}}
\end{aligned}
\tag{7}
$$

The suffix b stands for 'between FSU mean', W/w, 'within FSU population/ FSU sample and S/s stand for the population/ sample variance. A " ' " sign used for unequal SSU clusters.

3 Formulation of Proposed Estimator and Its Properties

We start with the basic Hansen Hurwitz mean estimator in a SRSWOR situation (i.e without considering cluster) which has the following form:

$$\bar{Y}_{HH} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}'_2 \tag{8}$$

Following (8), the one may modify [1] estimator to get its new version on the ith FSU while non-response exists even after second call, takes the form:

$$\hat{\bar{Y}}_{iHH} = \frac{m_{i1}}{m_i}\bar{y}_{i.1} + \frac{m_{i2}}{m_i}\bar{y}'_{i.2} \text{ where } \bar{y}'_{i.2} = \frac{1}{m_{i2r}}\sum_{j=1}^{m_{i2r}} y_{ij} \tag{9}$$

Finally, the [1] version of the estimator applicable for two-stage unequal size clusters takes the form:

$$T_{HH} = \frac{1}{n}\sum_{i=1}^{n}\hat{\bar{Y}}_{iHH} \tag{10}$$

In order to improve the second part of (9) we now wish to consider a combination of actual response and projected response from the final non-respondent ones through a difference type imputation procedure furnished below.

3.1 Description of Imputation Technique

The information on the second auxiliary variable x is readily available for all over the population U. Therefore, motivated by the imputation techniques suggested by [5] and [6], we propose the following difference type imputation method based on responding and non-responding units of the second stage sample S2after second call to estimate the responded part in equation (9) as described below:

We use the known responded part after second call in a difference type imputation scheme given below to estimate responded part $\bar{y}'_2$ by an estimator based on subsample of size m_{i2} per ith FSU supposing information on x is available including its population mean i.e. $\bar{X}_{..}$ and $\bar{X}_i$ .The details of imputation procedure (difference type) adopted has been shown below.

$$t_{ij} = \begin{cases} \frac{m_{i2}\bar{y}_{i2r}}{m_{i2r}} + b(\bar{X}_{i.} - x_{ij}) & j \in R \\ b(\bar{X}_{i.} - x_{ij}) & j \in R^c \end{cases} \tag{11}$$

where b is real constant to be determined suitably from past experiences.

The expression of estimator t_i (per FSU) is obtained as follows:

$$\begin{aligned}
t_{i2} &= \frac{1}{m_{i2}}\left(\sum_{j\in S_2} t_{ij}\right) = \frac{1}{m_{i2}}\left(\sum_{j\in R} t_i + \sum_{j\in R^c} t_{ij}\right) \\
&= \frac{1}{m_{i2}}\left[\sum_{j\in R}\frac{m_{i2}\bar{y}_{i2r}}{m_{i2r}} + b(\bar{X}_{i.} - x_{ij}) + \sum_{j\in R^c} b(\bar{X}_{i.} - x_{ij})\right] \\
&= \frac{1}{m_{i2}}\left[\frac{m_{i2}\bar{y}_{i2r}}{r} + \sum_{j\in R} b(\bar{X}_{i.} - x_{ij}) + \sum_{j\in R^c} b(\bar{X}_{i.} - x_{ij})\right] \\
&= \frac{1}{m_{i2}}\left[m_{i2}\bar{y}_{i2r} + bm_{i2r}\bar{X}_{i.} - \sum_{j\in R} bx_{ij} + b(m_{i2} - m_{i2r})\bar{X}_{i.} - \sum_{j\in R^c} bx_{ij}\right] \\
&= \frac{1}{m_{i2}}\left[m_{i2}\bar{y}_{i2r} + bm_{i2r}\bar{X}_{i.} - \sum_{j\in R} bx_{ij} - \sum_{j\in R^c} bx_{ij}\right] \\
&= \frac{1}{m_{i2}}\left[m_{i2}\bar{y}_{i2r} + bm_{i2r}\bar{X}_{i.} - b\left(\sum_{j\in R} x_{ij} + \sum_{j\in R^c} bx_{ij}\right)\right] \\
&= \frac{1}{m_{i2}}\left[m_{i2}\bar{y}_{i2r} + bm_{i2r}\bar{X}_{i.} - b\sum_{j\in S_2} x_{ij}\right] \\
&= \frac{1}{m_{i2}}\left[m_{i2}\bar{y}_{i.2r} + bm_{i2}\bar{X}_{i.} - bm_{i2}\bar{x}_{m_{i2}}\right] \\
&= \bar{y}_{i.2r} + b(\bar{X}_{i.} - \bar{x}_{m_{i2}})
\end{aligned}$$

$$\text{or } t_{i2} = \bar{y}_{i.2r} + b(\bar{X}_{i.} - \bar{x}_{m_{i2}}) = \bar{y}_{i.2r} + b(\bar{X}_{i.} - \bar{x}_{i.}) \tag{12}$$

We modify the second part $\frac{m_{i2}}{m_i}\bar{y}'_{i.2}$ of the ith $(i = 1,2,3,\cdot,N)$ cluster by replacing it with $\frac{m_{i2}}{m_i}\bar{t}_{i2}$ $(i = 1,2,3,\cdot,N)$ using imputation technique developed in equation (12) i.e., instead of taking a simple weighted average $\frac{m_{i2}}{m_i}\bar{y}'_{i.2}$ of responded part in second call we take a combination of actual response and projected response from the final non-respondents ones through the imputation procedure described above (equations (11) and (12)).

Therefore, the final proposed estimator T_p to estimate $\bar{Y}_{..}$ is defined as:

$$T_p = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m_i}\sum_{i=1}^{m_i} y_{ij}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m_i}\left(\sum_{i=1}^{m_i 1} y_{ij} + \sum_{i=1}^{m_i 2} y_{ij}\right)\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{m_{i1}}{m_i}\bar{y}_{i.1} + \frac{m_{i2}}{m_i}\bar{y}'_{i.2}\right] = \frac{1}{n}\sum_{i=1}^{n}\hat{\bar{Y}}_{HH}$$

$$\text{or } T_p = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{m_{i1}}{m_i}\bar{y}_{i.1} + \frac{m_{i2}}{m_i}[\bar{y}_{i.2r} + b(\bar{X}_{i.} - \bar{x}_{i.})]\right] \tag{13}$$

We assume here $\bar{X}_{....}, \bar{X}_{i.}$ and $\bar{x}_{i.2}$ are known at all levels.

4 Deriving the Expression of Variance from Proposed Estimator Structure

We have taken each FSU (of population size Mi where from a sample is drawn of size $m_i (i = 1,2,3,\cdots,N)$) as a dichotomous population in terms of respondents and non-respondents. Now we proceed to implement [1] estimator at each FSU level.

Finally, to obtain the mean square error of the proposed estimator presented in equation (13) for aforesaid situation, we proceed as follows.

It noted from equation (13) that

$$\text{or } T_p = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{m_{i1}}{m_i}\bar{y}_{i.1} + \frac{m_{i2}}{m_i}[\bar{y}_{i.2r} + b(\bar{X}_{i.} - \bar{x}_{i.})]\right]$$

We observe that

$$\bar{y}_{i.} = \frac{m_{i1}\bar{y}_{i.1} + m_{i2}\bar{y}'_{i.2}}{m_i} \tag{14}$$

$\bar{y}'_{i.2}$ is the sample mean of m_{i2} sample shown non-response at 1st call.

$$m_{i1}\bar{y}_{i.1} = \bar{y}_{i.}m_i - m_{i2}\bar{y}'_{i.2} \tag{15}$$

It is noted that $t_{i2} = \bar{y}_{i.2r} + b(\bar{X}_{i.} - \bar{x}_{i.1})$ from equation (12).

We replace $m_{i1}\bar{y}_{i.1}$ in equation (14) by expression in equation (15) and thus the estimator (12) becomes

$$T_p = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{m_i\bar{y}_{i.} - m_{i2}\bar{y}'_{i.1} + m_{i2}t_{i2}}{m_i}\right]$$

Hence the MSE(T_p) becomes

$E(T_p - \bar{Y}_{..})^2$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{m_i\bar{y}_{i.} - m_{i2}\bar{y}'_{i.2} + m_{i2}t_{i2}}{m_i} - \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_{i.}\right]^2$$
$$= E\left[\frac{1}{n}\left\{\sum_{i=1}^{n}\frac{m_i(\bar{y}_{i.} - \bar{Y}_{i.}}{m_i} + \sum_{i=1}^{n}\frac{m_{i2}(t_{i2} - \bar{y}'_{i.2})}{m_i}\right\}\right]^2$$
$$= E\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{y}_{i.} - \bar{Y}_{i.}) + \sum_{i=1}^{n}\frac{m_{i2}}{m_i}(t_{i2} - \bar{y}_{i.2})\right]^2$$

Expanding the above expression, we have the mean square error of the proposed estimator T_p as

$$V(T_p) = \begin{array}{l}\left(\frac{1}{n} - \frac{1}{N}\right)S'^2_{yb} + \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S'^2_{ywi} \\ +W_2(k-1)\left[S'^2_{y2b}(1-\rho^2_{yx2b})\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i} + \frac{1}{nN}\sum_{i=1}^{N}\frac{1}{m_i}S'^2_{y2wi}(1-\rho^2_{iyx2w})\right]\end{array} \tag{16}$$

where $S'^2_{yb} = \frac{1}{N-1}\sum_{i=1}^{n}(\bar{Y}_{i.} - \bar{Y}_{..})^2, S'^2_{wi} = \frac{1}{M_i - 1}\sum_{j=1}^{m_i}(Y_{ij} - \bar{Y}_{i.})^2.$

S'^2_{2yb} has the same structure as S'^2_{yb} but for non-response part. Also $\rho^2_{yx2b}, \rho^2_{yx2w}$ and S'^2_{2ywi} is for non-resp. part and same with S'^2_{wi}.

Similarly, from equation (10) the expression of variance of the [1] version of the estimator for two-stage unequal size cluster derived as

$$V(T_{HH}) = \left(\frac{1}{n} - \frac{1}{N}\right) S'^2_{yb} + \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S'^2_{ywi} + W_2(k-1)\left[S'^2_{y2b}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i} + \frac{1}{nN}\sum_{i=1}^{N}\frac{1}{m_i}S'^2_{y2wi}\right] \tag{17}$$

Remark 4.1: Readers are encouraged to see the demonstration given in Appendix about the procedures of obtaining variance of sample mean under unequal size clusters in presence of non-response situations.

Remark 4.2: One may consider $M_i = M, i = 1,2, \cdots, N; m_i = m, i = 1,2,\cdots,n$ for equal size clusters and thus derived the expressions for $V(T_p)$ and $V(T_{HH})$ from equations (16) and (17) as:

$$\begin{aligned} V(T_p) &= \left(\frac{1}{n} - \frac{1}{N}\right) S'^2_{yb} + \frac{1}{n(m-M)}\bar{S}'^2_{ywi} + W_2(k-1)\left[S'^2_{y2b}(1-\rho^2_{yx2b})\frac{1}{m} + \frac{1}{nmN}\sum_{i=1}^{N} S'^2_{y2wi}(1-\rho^2_{iyx2w})\right] \\ V(T_{HH}) &= \left(\frac{1}{n} - \frac{1}{N}\right) S'^2_{yb} + \frac{1}{n(m-M)}\bar{S}'^2_{ywi} + W_2(k-1)\left[S'^2_{y2b}\frac{1}{m} + \frac{1}{nm}\bar{S}'^2_{y2wi}\right] \end{aligned} \tag{18}$$

5 Theoretical Analysis of Efficiency

The difference of $V(T_{HH})$ and $V(T_p)$ from (16) and (17) as:

$$\begin{aligned} &V(T_{HH}) - V(T_p) = \Delta V \\ &= \left(\tfrac{1}{n} - \tfrac{1}{N}\right) S'^2_{yb} + \tfrac{1}{nN}\sum_{i=1}^{N}\left(\tfrac{1}{m_i} - \tfrac{1}{M_i}\right) S'^2_{ywi} + W_2(k-1)\left[S'^2_{y2b}\tfrac{1}{N}\sum_{i=1}^{N}\tfrac{1}{m_i} + \tfrac{1}{nN}\sum_{i=1}^{N}\tfrac{1}{m_i}S'^2_{y2wi}\right] - \\ &\left(\tfrac{1}{n} - \tfrac{1}{N}\right) S'^2_{yb} + \tfrac{1}{nN}\sum_{i=1}^{N}\left(\tfrac{1}{m_i} - \tfrac{1}{M_i}\right) S'^2_{ywi} + W_2(k-1)\left[S'^2_{y2b}(1-\rho^2_{yx2b})\tfrac{1}{N}\sum_{i=1}^{N}\tfrac{1}{m_i} + \tfrac{1}{nN}\sum_{i=1}^{N}\tfrac{1}{m_i}S'^2_{y2wi}(1-\rho^2_{iyx2w})\right] \\ &= W_2(k-1)\left[S'^2_{y2b}\rho^2_{yx2b}\tfrac{1}{N}\sum_{i=1}^{N}\tfrac{1}{m_i} + \tfrac{1}{nN}\sum_{i=1}^{N}\tfrac{1}{m_i}S'^2_{y2wi}\rho^2_{iyx2w}\right] > 0 \text{ always} \end{aligned} \tag{19}$$

This shows the proposed estimator T_p is always more precise than the [1] version estimator T_{HH} applicable for two-stage unequal size clusters in presence of non-response situation.

Remark 5.1: For equal SSU sizes at both population and sample level the difference (19) reduces to

$$\begin{aligned} &\Delta V_{eq} = \left(\tfrac{1}{n} - \tfrac{1}{N}\right) S'^2_{yb} + \tfrac{1}{n(m-M)}\bar{S}'^2_{ywi} + W_2(k-1)\left[S'^2_{y2b}\tfrac{1}{m} + \tfrac{1}{nm}\bar{S}'^2_{y2wi}\right] - \\ &\left(\tfrac{1}{n} - \tfrac{1}{N}\right) S'^2_{yb} - \tfrac{1}{n(m-M)}\bar{S}'^2_{ywi} - W_2(k-1)\left[S'^2_{y2b}(1-\rho^2_{yx2b})\tfrac{1}{m} + \tfrac{1}{nmN}\sum_{i=1}^{N}\bar{S}'^2_{y2wi}(1-\rho^2_{iyx2w})\right] \\ &= \tfrac{W_2(k-1)}{m}\left[S'^2_{y2b}\rho^2_{yxb} + \tfrac{1}{n}\bar{S}'^2_{y2w}\bar{\rho}^2_{yx2w}\right] \text{ where } \tfrac{1}{N}\sum_{1} NS'^2_{y2wi}\rho^2_{yxwi}) = \bar{S}'^2_{y2w}\bar{\rho}^2_{yx2w} \end{aligned} \tag{20}$$

Clearly $\Delta V_{eq} > 0$ also always

6 Empirical Investigation

As a mandatory step to test the efficacy of a proposed estimator, we have chosen two natural populations (described in Sec. 6.1.1 & 6.1.2) and an artificially generated one (Sec.6.2) and applied our estimator there. For examination of performance of the proposed strategy, we have examined the Percentage Relative Efficiency (PRE) of the estimator T_p with respect to the estimator T_{HH} as:

$$PRE = \frac{V(T_{HH})}{V(T_p)} \times 100 \tag{21}$$

The results are discussed below.

Table 1: (basic parameters computed based on data)

$\bar{Y}_{1.}$	$\bar{Y}_{2.}$	$\bar{Y}_{3.}$	$\bar{Y}_{4.}$	$\bar{Y}_{5.}$	$\bar{X}_{1.}$	$\bar{X}_{2.}$	$\bar{X}_{3.}$
102.36	107.63	90.11	89.40	108.63	885.93	552.46	564.18
$\bar{X}_{4.}$	$\bar{X}_{5.}$	$\bar{Y}_{..}$	$\bar{X}_{..}$	S'^2_{yb}	S'^2_{yw}	ρ^2_{yxB}	S'^2_{xb}
737.49	652.81	99.62	678.57	69.51	7326.99	0.00	15224.40

Table 2: PREs of T_P over T_{HH} for different choices of k

k	$V(T_p)$	$V(T_{HH})$	PRE
	SSU units: $m_1=6, m_2=4, m_3=4, m_4=9, m_5=4$		
3	323.37	784.51	242.60
2	293.77	524.34	178.48
	SSU units: $m_1=5, m_2=3, m_3=3, m_4=7, m_5=3$		
3	498.58	1097.53	220
2	459.44	758.92	165

Table 3: Formation of 5 different clusters (zone wise) out of 51 states of United States and corresponding parametric values (N=5, n=3 taken)

Zones (FSU s)	Constituent States(SSU s)	Statistical Parameters
FSU 1	Wyoming, Missouri, Mississippi, Kentucky, Oklahoma, Arkansas, Indiana, Nebraska, South Carolina, Wisconsin, Utah, South Dakota, Idaho, West Virginia	$M_1=14, m_1=8, \bar{X}_{1.}=6.551, \bar{Y}_{1.}=6.59$
FSU 2	Alaska, Montana, New Hampshire, Minnesota, Vermont , Ohio,Arizona,New Mexico,North Dakota, Maine, Michigan, Massachusetts, Washington	$M_2=13, m_2=8, \bar{X}_{2.}=15.031, \bar{Y}_{2.}=15.11$
FSU 3	Kansas, Virginia, North Carolina, Oregon, Pennsylvania, Texas, Louisiana, Colorado, Tennessee, Iowa, Alabama, Georgia	$M_3=12, m_3=6, \bar{X}_{3.}=15.031, \bar{Y}_{3.}=15.11$
FSU 4	Hawaii, Rhode, Island, Connecticut, Nevada, Florida, California,, Illinois	$M_4=7, m_4=5, \bar{X}_{4.}=24.562, \bar{Y}_{4.}=24.48$
FSU 5	Maryland, District of Columbia, New Jersey, New York, Delaware	$M_5=5, m_5=3, \bar{X}_{5.}=33.528, \bar{Y}_{5.}=33.55$

6.1 Natural population

6.1.1 Study Using Natural Data Set 1:

We have adopted the medical data (source Table 174. Community Hospitals–States: 2000 and 2009). For definition of community hospitals see footnote 2, Table 168]. The number of hospitals is taken as primary study variable Y and patients admitted (in 1000) as X. The data is clustered. The states are divided in 5 FSU s and within each FSU the hospitals are considered as SSU's.

We have taken $N=5, n=3, (1/n-1/N)=0.13, Let W_2=.5, k=3,2$(giving $W_2(k-1)=1, 0.5$) for all cluster since they are homogeneous across. The parameters calculated based on data shown in Table 3.

Finally we calculate the Variances and PRE (as per definition (21)) under different values of $k(=3,2)$ taking $W_2=0.5$ as furnished in Table 4.

6.1.2 Study Using Natural Data Set 2:

We have chosen another natural population datasets on abortion rates form Statistical Abstract of the United States: 2011 to elucidate the performance of our proposed estimator. The variables y and x denote number of abortions rates reported in the state of US during the years 2008 and 2007 respectively. Details about the parametric values illustrated in Table 5.

Table 4: Population mean squares and correlation coefficients of the respective variables

			S'^2_{ywi}	S'^2_{xwi}	ρ_{xywi}
FSU 1			4.56	4.51	0.9784
FSU 2			6.21	6.33	0.9013
FSU 3			5.87	5.66	0.9777
FSU 4			31.42	37.22	0.9885
FSU 5			19.31	29.03	0.9751
$\bar{X}$	$\bar{Y}$	S'^2_{yb}	S'^2_{yw}	S'^2_{xw}	ρ_{yxb}
18.94	18.97	23.25	13.47	16.45	0.04

Table 5: Variance and PRE of the proposed Estimator T_p with respect to T_{HH} obtained for Data Set 2 (taking $W_2 = .5$, k=3 and 2) and using PRE definition (21).

k	$V(T_p)$	$V(T_{HH})$	*PRE*
	SSU units: $m_1 = 6, m_2 = 4, m_3 = 4, m_4 = 9, m_5 = 4$		
3	16.52	19.31	116.88
2	15.05	18.04	119.17
	SSU units: $m_1 = 5, m_2 = 3, m_3 = 3, m_4 = 7, m_5 = 3$		
3	18.67	20.71	110.90
2	16.27	17.01	104.54

The computed values of population mean squares and correlation coefficients of the respective variables based on the ith FSU ($i = 1, 2, \cdots, 5$) are shown in Table 6:

To measure the performance of the proposed estimator T_p, we have computed the PREs of T_p with respect to T_{HH}. The findings are displayed in Table 5.

6.2 Study Using Artificially Generated Data Set Through Simulation:

Efficiency comparison through artificial population generation technique helps in concluding whether a newly developed technique is better than the existing ones (see for instance the work of [7]). In the context of statistical simulation is a computer driven program to create pseudo-random sampling data. A key strength of simulation study is the ability to analyze the behavior of statistical methods (i.e. forming and testing a new estimator etc.) from some parameter(s) of interest that can be reliably generated from the simulation process.

In what follows we have rigorously investigated the efficacy of our estimator by simulating a normal population of moderately large sample size. We have largely adopted the techniques followed [8,11,9,10,7].

6.2.1 Population Generated by Simulation Using Standard Normal Distribution

Probability density function is X. (continuous) We have generated three sets of independent random numbers of size N (N = 100) namely $x1[k], y1[k] (k = 1, 2, \cdots, N)$ from a standard normal distribution as presented below.

The following algorithm is used to generate the population artificially:

1.Generate two random variables which are normally distributed with mean 0, S.D. =1 and are of size 100; (Note: are temporary variables).
2.Define N =100
3.Define rx1y1 = 0.5
Define correlation coefficient ($r_{x1y1} = 0.5$) and standard deviations of the variables y1 and x1 as $SX1 = \sqrt{50}$, $SY1 = \sqrt{40}$ and their mean as $m_{x1} = 20, m_{y1} = 25$ respectively.
4.a=sy1*sy1*(1-(rx1y1^2))

Table 6: PRE of the proposed Estimator with respect to estimator obtained for Artificially Generated Population Data Set 1 (taking $W_2 = .5, k = 3, 2, n = 3$ and $m_i = 10, 13, 15$) and using PRE definition (21).

	$V(T_p)$	$V(T_{HH})$	$PRE(T_P$ wrt $T_{HH})$
Correlation(x, y) = 0.5, n= 3, mi= 10for all i			
3	155.94	162.23	104.03
2	148.02	154.76	104.55
Correlation(x, y) = 0.5, n=3, $m_i = 13$ for all i			
3	157.55	159.48	101.23
2	149.97	151.09	100.75
Correlation(x, y) = 0.5, n= 3, $m_i = 15$ for all i			
3	171.25	192.19	112.23
2	166.34	176.56	106.75
Correlation(x, y) = 0.75, n= 3, $m_i = 10$ for all i			
3	145.24	149.94	103.23
2	136.98	143.02	104.41
Correlation(x, y) = 0.75, n= 3, $m_i = 13$ for all i			
3	147.24	155.91	105.89
2	134.55	149.68	111.25
Correlation(x, y) = 0.75, n= 3, $m_i = 15$ for all i			
3	151.74	164.18	108.20
2	145.64	174.49	119.81

5.for(j in 1:N)
{
x[j]= +(sx1x1[j])
y[j]= +($\sqrt{a}$)y1[j])+(rx1y1*sy1x1[j])
}
6.Take output of the variables x and y.
7.Repeat the steps 1 to 6 with different values of rx1y1 (step 3) which will generate different population for different values of the correlation coefficients.

After generating the population artificially, we have taken 3 (i.e., n = 3) out of 5 clusters each of equal size of 10 ($M_i = 20, m_i = 10, i = 1, 2, 3$) from a total of 100 sequentially and computed the PRE of the proposed estimator T_p with respect to T_{HH} for different values of correlation coefficient ryx. The findings are displayed in Table 6.

7 Conclusion

Following outcomes may be observed from the above study:

1.From equation (16), (17) and (18) it is found theoretically that in case of equal or unequal size clusters, our suggested methodology of utilizing imputation instead of taking simply weighted average of responded part after second call always gives an edge.
2.From tables 4 and 7 corresponding to the natural population data set 1 and 2 respectively, it is observed that the efficiency is increased with larger SSU sizes (m_is). It is also found that, we have more efficiency for our proposed strategy even when the sub sampling fractions (i.e., k) is larger for which we require to sample only smaller number of units at second calls. This establishes that that imputation is definitely having some positive effect on the survey and also reduces the cost of the surveys. It may be noted that natural population data set is heterogeneous because the parameters vary significantly from cluster to cluster. Hence, it is established that our proposed methodology performs profoundly in practical surveys where population is often heterogeneous as well as cluster sizes may vary.
3.From table 8 corresponding to the data set in artificially generated population through simulation studies, it may be seen that for increasing choices of correlation coefficients, we have better performance from our proposed methodology. These findings lead us in choosing the suitable population where our suggested strategy may be applicable. It also noted that artificially generated population is homogeneous because parametric values do not vary significantly from cluster to clusters and our suggested strategy performs effectively there in comparison to the conventional approach.

4.From tables 4, 7 and 8, it is found that for different choices of sample SSUs from each cluster (for example, mi for ith cluster), our proposed estimator produces more efficiency than the conventional approach.

Therefore, the proposed methodology may be found effective in comparison with the contemporary ones as it unifies several merits in one structure. Hence, looking on the encouraging outcomes, we are happy to recommend our proposed strategy to the survey statisticians in practice.

References

[1] M. H. Hansen, W. N. Hurwitz, The problem of non-response in sample surveys. Journal of the American Statistical Association, **41**, 517-529.
[2] I. G. Sande , A personal view of hot-deck imputation procedures. Survey Methodology, **5**, 238-247 (1979).
[3] L. Chand, Some ratio-type estimators based on two or more auxiliary variable. Unpublished Ph. D. thesis.Iowa State University, Ames, Iowa (USA), (1975).
[4] H. Lee, E. Rancourt, C. E. Sarndall (1994): Experiments with variance estimation from survey data with imputed values. Journal of Official Statistics, **10(3)**, 231-243 (1946).
[5] S. Singh , A new method of imputation in survey sampling. Statistics, **43(5)**, 499–511 (2009).
[6] G. Diana, P. F. Perri, P. F., Improved estimators of the population mean for missing data. Communications in Statistics –Theory and Methods,**39**, 3245–3251 (2010).
[7] G. N. Singh, A. K. Sharma, A. Bandyopadhyay, Effectual variance estimation strategy in two occasions successive sampling in presence of random non response. Communications in Statistics - Theory and Methods (2016), DOI.10.1080/03610926.2016.1146769.
[8] S. Singh, B. Deo , Imputation by power transformation. Statistical Papers, **4**, 555-579 (2003).
[9] G. N. Singh, A. K. Sharma, A. Bandyopadhyay, Effectual Variance Estimation Strategy in Two Occasions Successive Sampling in Presence of Random Non-Response. Communications in Statistics-Theory & Methods, (2017), DOI:10.1080/03610926.2016.1146769.
[10] S. Maurya, M. Khetan, C. Cadilar, G. N. Singh, Some imputation methods for missing data in sample surveys. Hacettepe University Bulletin of Natural Sciences and Engineering- Series B. Mathematics and Statistics (2015), DOI.10.15672/HJMS.20159714095.
[11] G. N. Singh, M. Khalid,Some imputation methods to compensate with non-response for estimation of population mean in two occasion successive sampling. Communications in Statistics-Theory and Methods, **49(14)**, 3329-3351 (2020).

Appendix:

Deriving variance expression for total and mean in 2-Stage unequal cluster population

$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} M_i\bar{y}_{i.}$ (estimates FSU total)

$\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i.}$ (estimates FSU mean)

which are unbiased. $M_0 = \sum_{i|1}^{N} M_i =$ total SSU units.

for equal SSU size these are $\hat{Y} = MN\left[\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}\right]$

$\hat{\bar{Y}}_{fsu} = \frac{M}{n}\sum_{i=1}^{n}\bar{y}_{i.}$ and $\hat{\bar{Y}}_{..} = \frac{\hat{Y}}{M_0} = \frac{\hat{Y}/N}{M} = \frac{\hat{\bar{Y}}_{fsu}}{M}$ where $M_0 = MN$ ($\hat{\bar{Y}}$ actually estimates $\bar{Y}_{ij}$)

$V(\hat{Y}) = V_1E_2(\hat{Y}) + E_1V_2(\hat{Y})$ The first term is for variance between FSU's and the 2nd term is for within SSU per FSU (E_1, V_1 stand for expectation, variance for all possible selections of FSU and E_2, V_2 stand for expectation, variance for all possible selections of SSU within a FSU).

Then

$E_1V_2(\hat{Y}) = E_1\left(\frac{N^2}{n^2}\sum_{i=1}^{n} M_i^2V_2(\bar{y}_{i.})\right)$; $V_2(\bar{y}_{i.})$ is the sample variance of ith FSU

$= E_1\left(\frac{N^2}{n^2}\sum_{i=1}^{n}\frac{M_i^2}{m_i}(1-f_{2i})s_{i.}^2\right) = \left(\frac{N^2}{n^2}\frac{n}{N}\sum_{i=1}^{n}\frac{M_i^2}{m_i}(1-f_{2i})S_{i.}^2\right)$

where $f_{2i} = \frac{m_i}{M_i}$ and $S_{i.}^2 = \sum_{i=1}^{m_i}\frac{Y_{ij}-\bar{Y}_{i.}^2}{M_i-1}$

Similarly

$V_1E_2(\hat{Y}) = V_1E_2(\frac{N}{n}\sum_{i=1}^{n}M_i\bar{y}_{i.}) = V_1(\frac{N}{n}\sum_{i=1}^{n}Y_{i.}) = \frac{N^2}{n^2}\sum_{i=1}^{n}V_1(Y_{i.})$; since $M_i\bar{Y}_{i.} = Y_i$ (FSU total)

$= \frac{N^2}{n^2}\frac{n(1-\frac{1}{n})}{N-1}\sum_{i=1}^{N}(Y_{i.}-\bar{Y}_{i.})^2 = \frac{N^2}{n}(1-f_1)\bar{M}^2S_b'^2 = \frac{N^2}{n}(1-f_1)S_B'^2$

where $S_B'^2 = \sum_{i=1}^{N}\frac{(\frac{M_iY_i}{M}-\bar{Y})^2}{N-1}$, population variance between FSU total.

Thus the population total estimator $V(\hat{Y}) = \frac{N^2}{n^2}\frac{n}{N}\sum_{i=1}^{N}\frac{M_i^2}{m_i}(1-f_{2i})S_{i.}^2 + \frac{N^2}{n}(1-f_1)S_B'^2$

Similarly we obtain the population mean estimator as follows

$E_1V_2(\hat{\bar{Y}} = E_1\left(\frac{1}{n^2}\sum_{i=1}^{n}V_2(\bar{y}_{i.})\right) = \frac{1}{n^2}\frac{n}{N}\sum_{i=1}^{N}\frac{1}{m_i}(1-f_{2i})S_{i.}^2 = \frac{1}{nN}\sum_{i=1}^{N}\frac{1}{m_i}(1-f_{2i})S_{i.}^2$

$V_1E_2(\hat{\bar{Y}})\frac{1}{n^2}\sum_{i=1}^{n}V_1(Y_{i.}) = \frac{1}{n}(1-f_1)S_b'^2, f_1 = \frac{n}{N}S_b'^2$ is the population variance between FSU mean.

Hence finally

$V(\hat{\bar{Y}}) = \frac{1}{nN}\sum_{i=1}^{N}\frac{1}{m_i}(1-f_{2i})S_{i.}^2 + \frac{1}{n}(1-f_1)S_b'^2$ where $S_b'^2 = \sum_{i=1}^{N}\frac{(\frac{M_i\bar{Y}_{i.}}{M}-\bar{Y}_{..}^2)^2}{N-1}$