Research Journal of Mathematical and Statistical Sciences ______________________________ISSN 2320–6047
Vol. **1(3),** 1-9, April **(2013)** *Res. J. Mathematical & Statistical Sci.*

On Mean Estimation with Imputation in Two- Phase Sampling Design

Thakur Narendra Singh[1], Yadav Kalpana[2] and Pathak Sharad[3]
[1]Center for Mathematical Sciences (CMS), Banasthali University, Rajasthan, INDIA
[2]Department of Statistics, University of Delhi, Delhi, INDIA
[3]Department of Mathematics and Statistics, Dr. H. S. Gour Central University, Sagar MP, INDIA

Available online at: www.isca.in
Received 02[nd] July 2012, revised 16[th] February 2013, accepted 19[th] February 2013

Abstract

A sample survey remains incomplete in presence of missing data and one of the substitution techniques of missing observations is known as imputation. A number of imputation methods are available in literature using auxiliary information, for example, Mean method of imputation, Ratio method of imputation, Compromised method of imputation and so on. These suggested methods are based on either population parameter of auxiliary variable or available information (both study and auxiliary variable) in the sample. Also, the number of available observations is considered as a constant but practically, it is not possible, the missing values may vary from sample to sample i.e. it may be considered as random variable. If population mean of auxiliary variable is unknown, then all these methods fail to perform. In such situations the idea of two-phase sampling is used for estimating population parameters. This paper presents the estimation of mean in presence of missing data under two-phase sampling scheme while the numbers of available observations are considered as random variable. The bias and m.s.e of suggested estimators are derived in the form of population parameters using the concept of large sample approximation. Numerical study is performed over two populations using the expressions of bias and m.s.e and efficiency compared with existing estimators.

Keywords: Estimation, missing data, imputation, post-stratification, bias, mean squared error (M.S.E.).

Introduction

Missing data is a problem encountered in almost every data collection activity but particularly in sample survey. The missing data naturally occurs in sample surveys when some, not all sampling units refuse or unable to participate in the survey or when data for specific items on a questionnaire completed for an otherwise cooperating unit are missing. For this, some imputation techniques are derived in literature by many authors to replace the missing part. Imputation is a methodology, which uses available data as a tool for the replacement of missing observations.

In literature, several imputation techniques are described, some of them are better over others. There is three concepts advocated by authors for missingness pattern: OAR (observed at random), MAR (missing at random), and PD (parametric distribution)[1]. If the probability of the observed missingness pattern, given the observed and unobserved data, does not depend on the value of the unobserved data, then data are missing at random (MAR). The observed data are observed at random (OAR) if for each possible value of the missing data and the parameter ϕ the conditional probability of the observed pattern of missing data given the missing data and the observed data, is the same for all possible values of the observed data.

There are different ways and means to control non-response. One way of dealing with the problem of non-response is to make more efforts to collect information by taking a sub-sample of units not responding at the first attempt. Another way of dealing with the problem of non-response is to estimate the probability of responding informants of their being at home at a specified point of time and weighting results with the inverse of this probability. A technique to deal with the problem of non-response was developed under the assumption that the population is divided into two classes, a *response class* who respond in the first attempt and a *non-response class* who did not.[2]

A questionnaire contains many questions that we call items. When item non-response occurs, substantial information about the non-respondent is usually available from other items on the questionnaire. Many imputation methods in literature use selection of these items as auxiliary variable in assigning values to the i^{th} non-respondent for item *y*.

Let the variable Y is of main interest and X be an auxiliary variable correlated with Y and the population mean $\overline{X}$ of auxiliary variable is unknown. A large preliminary simple random sample (without replacement) $S^{'}$ of $n^{'}$ units is drawn from the population $\Omega = \{1, 2, ..., N\}$ to estimate $\overline{X}$ and a secondary sample *S* of size n ($n < n'$) drawn as a sub-sample of the sample S' to estimate the population mean of main variable. Let the sample *S* contains n_1 responding units and $n_2 = (n - n_1)$ non-responding units. Using the concept of post-stratification, sample may be divided into two groups: responding (R_1) and non-responding (R_2).

The sample may be considered as stratified into two classes namely a *response class* and *non-response class,* and then the procedure is known as *post-stratification*. Post-stratification procedure is as precise as the stratified sampling under proportional allocation if the sample size is large enough[3]. Estimation problem in sample surveys, in the setup of post-stratification, under non-response situation is studied and given the concept of utilization of available information related to auxiliary variable *X* in imputation for missing observations of auxiliary information due to non-response[4,5,6].

Now it may be consider the population has two types of individuals like N_1 as number of respondents (R_1) and N_2 non-respondents (R_2), Thus the total *N* units of the population will comprise N_1 and N_2, respectively, such that $N = N_1+N_2$. The population proportions of units in the R_1 and R_2 groups are expressed as $W_1 = N_1/N$ and $W_2 = N_2/N$ such that $W_1+W_2=1$. Further, let $\overline{Y}$ and $\overline{X}$ be the population means of *Y* and *X* respectively. For every unit $i \in R_1$, the value y_i is observed available. However, for the units $i \in R_2$, the y_i's are missing and imputed values are to be derived. The i^{th} value x_i of auxiliary variate is used as a source of imputation for missing data when $i \in R_2$. This is to assume that for sample *S*, the data $x_s = \{x_i : i \in S\}$ are known. The following notations are used in the present research manuscript:

$\bar{x}_n, \bar{y}_n$: the sample mean of X and Y respectively in *S*; $\bar{x}_1, \bar{y}_1$: the sample mean of X and Y respectively in R_1 ; S_X^2, S_Y^2: the population mean squares of X and Y respectively; C_X, C_Y: the coefficient of variation of X and Y
ρ Correlation Coefficient in population between *X* and *Y*.

Further, consider few more symbolic representations: $L = E\left(\frac{1}{n_1}\right) = \left[\frac{1}{nW_1} + \frac{(N-n)(1-W_1)}{(N-1)n^2W_1^2}\right]$, $M = \frac{(N-n)(n-n_1)n'N}{nn_1^2(N-1)(N-n')}$,

Large Sample Approximation

Let $\bar{y}_1 = \overline{Y}(1+e_1)$; $\bar{x}_1 = \overline{X}(1+e_2)$; $\bar{x}_n = \overline{X}(1+e_3)$ and $\bar{x}' = \overline{X}(1+e_3')$, which implies the results $e_1 = \frac{\bar{y}_1}{\overline{Y}} - 1$; $e_2 = \frac{\bar{x}_1}{\overline{X}} - 1$;

$e_3 = \frac{\bar{x}_n}{\overline{X}} - 1$ and $e_3' = \frac{\bar{x}'}{\overline{X}} - 1$. Now by using the concept of two-phase sampling and the mechanism of MCAR,[7] for given n_1, n and n' we have :

$$E(e_1) = E[E(e_1)|n_1] = E\left[\left(\frac{\bar{y}_1 - \overline{Y}}{\overline{Y}}\right)\Bigg|n_1\right] = \frac{\overline{Y} - \overline{Y}}{\overline{Y}} = 0 \text{; Similarly, } E(e_2) = E(e_3) = E(e_3') = 0$$

$$E(e_1^2) = E\left[\left(\frac{\bar{y}_1 - \overline{Y}}{\overline{Y}}\right)^2\Bigg|n_1\right] = \left(E\left(\frac{1}{n_1}\right) - \frac{1}{n'}\right)C_y^2 = \left(L - \frac{1}{n'}\right)C_y^2 \text{ Similarly, } E(e_2^2) = \left(L - \frac{1}{n'}\right)C_X^2 \text{ ; } E(e_3^2) = \left(\frac{1}{n} - \frac{1}{n'}\right)C_X^2 \text{;}$$

$$E(e'^2_3) = \left(\frac{1}{n'} - \frac{1}{N}\right)C_X^2 \text{; } E(e_1e_2) = E(e_1e_2/n_1) = E\left[\left(\frac{(\bar{y}_1 - \overline{Y})(\bar{x}_1 - \overline{X})}{\overline{Y}\,\overline{X}}\right)\Bigg|n_1\right] = \left(E\left(\frac{1}{n_1}\right) - \frac{1}{n'}\right)\rho C_y C_X = \left(L - \frac{1}{n'}\right)\rho C_y C_x$$

$$E(e_1e_3) = \left(\frac{1}{n} - \frac{1}{n'}\right)\rho C_y C_X \quad E(e_1e_3') = \left(\frac{1}{n'} - \frac{1}{N}\right)\rho C_y C_X \text{; } E(e_2e_3) = \left(\frac{1}{n} - \frac{1}{n'}\right)C_X^2 \text{;}$$

$$E(e_2e_3') = \left(\frac{1}{n'} - \frac{1}{N}\right)C_X^2 \text{; } E(e_3e_3') = \left(\frac{1}{n'} - \frac{1}{N}\right)C_X^2 \text{;}$$

Some Existing Imputation Methods

Let $\overline{Y} = N^{-1}\sum_{i=1}^{N} y_i$ be the mean of the finite population under consideration. A Simple Random Sampling Without Replacement (SRSWOR), S, of size n is drawn form $\Omega = \{1,2,...N\}$ to estimate the population mean $\overline{Y}$. Let the number of responding units out of sampled n units be denoted by $r(r<n)$, the set of responding units, by R, and that of non- responding units by R^c. For every unit $i \in R$ the value y_i is observed, but for the units $i \in R^C$, the observations y_i are missing and instead imputed values are derived. The i^{th} value x_i of auxiliary variate is used as a source of imputation for missing data when $i \in R^C$. Assume for S, the data $x_s = \{x_i : i \in S\}$ are known with mean $\overline{x} = (n)^{-1}\sum_{i'=1}^{n'} x_{i'}$ .Under this setup, some well known imputation methods are given below:

Mean Method of Imputation:

For y_i define $y_{\bullet i}$ as $$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \overline{y}_r & if \quad i \in R^C \end{cases} \tag{1}$$

Using above, the imputation-based estimator of population mean $\overline{Y}$ is: $$\overline{y}_m = \frac{1}{r}\sum_{i \in R} y_i = \overline{y}_r \tag{2}$$

The bias and mean squared error is given by

(i) $$B(\overline{y}_m) = 0 \tag{3}$$

(ii) $$V(\overline{y}_m) \approx \left(\frac{1}{r} - \frac{1}{N}\right)S_y^2 \tag{4}$$

Ratio Method of Imputation:

For y_i and x_i, define $y_{\bullet i}$ as $$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \hat{b}x_i & if \quad i \in R^C \end{cases} \quad \text{where } \hat{b} = \sum_{i \in R} y_i \Big/ \sum_{i \in R} x_i \tag{5}$$

Using above, the imputation-based estimator is: $$\overline{y}_S = \frac{1}{n}\sum_{i \in S} y_{\bullet i} = \overline{y}_r\left(\frac{\overline{x}_n}{\overline{x}_r}\right) = \overline{y}_{RAT} \tag{6}$$

where $\overline{y}_r = \frac{1}{r}\sum_{i \in R} y_i$, $\overline{x}_r = \frac{1}{r}\sum_{i \in R} x_i$ and $\overline{x}_n = \frac{1}{n}\sum_{i \in S} x_i$

Lemma: The bias and mean squared error of $\overline{y}_{RAT}$ is given by

(i) $$B(\overline{y}_{RAT}) = \overline{Y}\left(\frac{1}{r} - \frac{1}{n}\right)\left(C_x^2 - \rho C_y C_x\right) \tag{7}$$

(ii) $$M(\overline{y}_{RAT}) \approx \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\left[S_y^2 + R_1^2 S_x^2 - 2R_1 S_{xy}\right] \quad \text{where } R_1 = \frac{\overline{Y}}{\overline{X}} \tag{8}$$

Compromised Method of Imputation:[8]

$$y_{\bullet i} = \begin{cases} (\alpha n/r)y_i + (1-\alpha)\hat{b}x_i & if \quad i \in R \\ (1-\alpha)\hat{b}x_i & if \quad i \in R^C \end{cases} \tag{9}$$

where α is a suitably chosen constant, such that the resultant variance of the estimator is optimum. The imputation-based estimator, for this case, is

$$\bar{y}_{COMP} = \left[\alpha \bar{y}_r + (1-\alpha)\bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}\right] \tag{10}$$

Lemma: The bias, mean squared error and minimum mean squared error at $\alpha = 1 - \rho \dfrac{C_Y}{C_X}$ of $\bar{y}_{COMP}$ is given by

(i) $B(\bar{y}_{COMP}) = \bar{Y}(1-\alpha)\left(\frac{1}{r}-\frac{1}{n}\right)\left(C_X^2 - \rho C_Y C_X\right)$ (11)

(ii) $M(\bar{y}_{COMP}) \approx \left\{\left(\frac{1}{n}-\frac{1}{N}\right)S_y^2 + \left(\frac{1}{r}-\frac{1}{n}\right)\left[S_y^2 + R_1^2 S_z^2 - 2R_1 S_{xy}\right]\right\} - \left(\frac{1}{r}-\frac{1}{n}\right)\alpha^2 \bar{Y}^2 C_x^2$ (12)

(iii) $M(\bar{y}_{COM})_{min} = \left[\left(\frac{1}{r}-\frac{1}{N}\right) - \left(\frac{1}{r}-\frac{1}{n}\right)\rho^2\right] S_Y^2$ (13)

Ahmed Methods:[9]

For the case where y_{ji} denotes the i^{th} available observation for the j^{th} imputation method:

(1) $$y_{1i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)}\left[n\bar{y}_r\left(\dfrac{\bar{X}}{\bar{x}_n}\right)^{\beta_1} - r\bar{y}_r\right] & if \quad i \in R^C \end{cases} \tag{14}$$

Under this, point estimator is $t_1 = \bar{y}_r\left(\dfrac{\bar{X}}{\bar{x}_n}\right)^{\beta_1}$ (15)

Lemma 1: The bias, mean squared error and minimum mean squared error at $\beta_1 = \rho \dfrac{C_Y}{C_X}$ of t_1 is given by

(i) $B(t_1)_I = \bar{Y}\left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{\beta_1(\beta_1+1)}{2}C_X^2 - \beta_1 \rho C_Y C_X\right)$ (16)

(ii) $M(t_1)_I \approx \bar{Y}^2\left[\left(\frac{1}{r}-\frac{1}{N}\right)C_Y^2 + \beta_1^2\left(\frac{1}{n}-\frac{1}{N}\right)C_X^2 - 2\beta_1\left(\frac{1}{n}-\frac{1}{N}\right)\rho C_Y C_X\right]$ (17)

(iii) $M(t_1)_{min} \approx \left(\frac{1}{r}-\frac{1}{N}\right)S_Y^2 - \left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{XY}^2}{S_X^2}$ (18)

(2) $$y_{2i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)}\left[n\bar{y}_r\left(\dfrac{\bar{x}_n}{\bar{x}_r}\right)^{\beta_2} - r\bar{y}_r\right] & if \quad i \in R^C \end{cases} \tag{19}$$

The point estimator is $t_2 = \bar{y}_r\left(\dfrac{\bar{x}_n}{\bar{x}_r}\right)^{\beta_2}$ (20)

Lemma: The bias, mean squared error and minimum mean squared error at $\beta_2 = \rho \dfrac{C_Y}{C_X}$ of t_2 is given by

(i) $B(t_2) = \left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}\left(\frac{\beta_2(\beta_2+1)}{2}C_X^2 - \beta_2 \rho C_Y C_X\right)$ (21)

(ii) $M(t_2)_I \approx \overline{Y}^2\left[\left(\frac{1}{r}-\frac{1}{N}\right)C_Y^2+\beta_2^2\left(\frac{1}{r}-\frac{1}{n}\right)C_X^2-2\beta_2\left(\frac{1}{r}-\frac{1}{n}\right)\rho C_Y C_X\right]$ (22)

(iii) $M(t_2)_{min} \approx \left(\frac{1}{r}-\frac{1}{N}\right)S_Y^2-\left(\frac{1}{r}-\frac{1}{n}\right)\frac{S_{XY}^2}{S_X^2}$ (23)

(3) $$y_{3i}=\begin{cases} y_i & if \quad i\in R \\ \frac{1}{(n-r)}\left[n\,\overline{y}_r\left(\frac{\overline{X}}{\overline{x}_n}\right)^{\beta_3}-r\,\overline{y}_r\right] & if \quad i\in R^C \end{cases} \tag{24}$$

The point estimator is $t_3=\overline{y}_r\left(\frac{\overline{X}}{\overline{x}_r}\right)^{\beta_3}$ (25)

Lemma: The bias, mean squared error and minimum mean squared error at $\beta_3=\rho\frac{C_Y}{C_X}$ of t_3 is given by

(i) $B(t_3)=\left(\frac{1}{r}-\frac{1}{N}\right)\overline{Y}\left(\frac{\beta_3(\beta_3+1)}{2}C_X^2-\beta_3\rho C_Y C_X\right)$ (26)

(ii) $M(t_3) \approx \left(\frac{1}{r}-\frac{1}{N}\right)\overline{Y}^2\left[C_Y^2+\beta_3^2C_X^2-2\beta_3\rho C_Y C_X\right]$ (27)

(iii) $M(t_3)_{min} \approx \left(\frac{1}{r}-\frac{1}{N}\right)S_Y^2(1-\rho^2)$ (28)

and when $k=4,\beta_l=0$ *then* $T_{FTl}=t_l=\overline{y}_r;(l=1,2,3)$

Proposed Different Imputation Methods

Let y_{vji} denotes the i^{th} available observation for the j^{th} imputation. We suggest the following imputation methods:

(4) $$y_{V1i}=\begin{cases} y_i & if \quad i\in R_1 \\ \frac{\overline{y}_1}{(1-W_1)}\left[\left(\frac{\overline{x}^{\cdot}}{\overline{x}_n}\right)^{\alpha}-W_1\right] & if \quad i\in R_2 \end{cases} \tag{29}$$

where α is suitably chosen constant, such that the variance the resultant estimator is minimum. Under this method, the point estimator of $\overline{Y}$ is $T_{V1}=\overline{y}_1\left(\frac{\overline{x}^{\cdot}}{\overline{x}_n}\right)^{\alpha}$ (30)

(5) $$y_{V2i}=\begin{cases} y_i & if \quad i\in R_1 \\ \frac{\overline{y}_1}{(1-W_1)}\left[\left(\frac{\overline{x}_n}{\overline{x}_1}\right)^{\beta}-W_1\right] & if \quad i\in R_2 \end{cases} \tag{31}$$

where β is suitably chosen constant, such that the variance the resultant estimator is minimum. Under this method, the point estimator of $\overline{Y}$ is $T_{V2}=\overline{y}_1\left(\frac{\overline{x}_n}{\overline{x}_1}\right)^{\beta}$ (32)

(6) $$y_{V3i} = \begin{cases} y_i & \text{if} \quad i \in R_1 \\ \dfrac{\bar{y}_1}{(1-W_1)}\left[\left(\dfrac{\bar{x}'}{\bar{x}_1}\right)^{\gamma} - W_1\right] & \text{if} \quad i \in R_2 \end{cases} \tag{33}$$

Where γ is suitably chosen constant, such that the variance the resultant estimator is minimum. Under this method, the point estimator of $\bar{Y}$ is $$T_{V3} = \bar{y}_1\left(\frac{\bar{x}'}{\bar{x}_1}\right)^{\gamma} \tag{34}$$

Note: At $\gamma = 1\ (-1)$ then the estimator T_{V3} convert into ratio (product) type estimator in two-phase sampling scheme.

Bias and MSE of Proposed Methods

Let $B(.)$ and $M(.)$ denote the bias and mean squared error (*M.S.E.*) of an estimator under a given sampling design. The properties of estimators are derived in the following theorems respectively.

Theorem 1:

(1) Estimator T_{V1} in terms of e_i; $i = 1,2,3$ and $e_3^{'}$, could be expressed upto first order of approximation:

$$T_{V1} = \bar{Y}\left[1 + e_1 + \alpha\left\{e_3^{'} - e_3 - e_1e_3 + e_1e_3^{'} - \alpha e_3e_3^{'} + \frac{\alpha+1}{2}e_3^2 + \frac{\alpha-1}{2}e_3^{'2}\right\}\right] \tag{35}$$

Proof: $$T_{V1} = \bar{y}_1\left(\frac{\bar{x}'}{\bar{x}_n}\right)^{\alpha} = \bar{Y}(1+e_1)(1+e_3^{'})^{\alpha}(1+e_3)^{-\alpha}$$

$$= \bar{Y}(1+e_1)\left(1 + \alpha e_3^{'} + \frac{\alpha(\alpha-1)}{2}e_3^{'2}\right)\left(1 - \alpha e_3 + \frac{\alpha(\alpha+1)}{2}e_3^2\right)$$

$$= \bar{Y}\left[1 + e_1 + \alpha\left\{e_3^{'} - e_3 - e_1e_3 + e_1e_3^{'} - \alpha e_3e_3^{'} + \frac{\alpha+1}{2}e_3^2 + \frac{\alpha-1}{2}e_3^{'2}\right\}\right]$$

(2) Bias of T_{V1} is: $$B(T_{V1}) = \bar{Y}\alpha\left(\frac{1}{n} - \frac{2}{n'} + \frac{1}{N}\right)\left(\frac{(\alpha+1)}{2}C_X^2 - \rho C_Y C_X\right) \tag{36}$$

Proof: $$B(T_{V1}) = E\left[T_{V1} - \bar{Y}\right] = \bar{Y}E\left[1 + e_1 + \alpha\left\{e_3^{'} - e_3 - e_1e_3 + e_1e_3^{'} - \alpha e_3e_3^{'} + \frac{\alpha+1}{2}e_3^2 + \frac{\alpha-1}{2}e_3^{'2}\right\}\right]$$

$$= \bar{Y}\alpha\left(\frac{1}{n} - \frac{2}{n'} + \frac{1}{N}\right)\left(\frac{(\alpha+1)}{2}C_X^2 - \rho C_Y C_X\right)$$

(3) Mean squared error of T_{V1}, upto first order of approximation could be written as:

$$M(T_{V1}) = \bar{Y}^2\left[\left(L - \frac{1}{n'}\right)C_Y^2 + \left(\frac{1}{n} - \frac{2}{n'} + \frac{1}{N}\right)\left(\alpha^2C_X^2 - 2\alpha\rho C_Y C_X\right)\right] \tag{37}$$

Proof: $$M(T_{V1}) = E\left[T_{V1} - \bar{Y}\right]^2$$

$$= \bar{Y}^2 E\left[1 + e_1 + \alpha\left\{e_3^{'} - e_3 - e_1e_3 + e_1e_3^{'} - \alpha e_3e_3^{'} + \frac{\alpha+1}{2}e_3^2 + \frac{\alpha-1}{2}e_3^{'2}\right\}\right]^2 = \bar{Y}^2\left[\left(L - \frac{1}{n'}\right)C_Y^2 + \left(\frac{1}{n} - \frac{2}{n'} + \frac{1}{N}\right)\left(\alpha^2C_X^2 - 2\alpha\rho C_Y C_X\right)\right]$$

(4) Minimum mean squared error of T_{V1} is:

$$\left[M(T_{V1})\right]_{Min} = \left[\left(L - \frac{1}{n'}\right) - \left(\frac{1}{n} - \frac{2}{n'} + \frac{1}{N}\right)\rho^2\right]S_Y^2 \text{ when } \alpha = \rho\frac{C_Y}{C_X} \tag{38}$$

Proof: First differentiate (38) with respect to α and then equate to zero, we get

$$\frac{d}{d\alpha}\left[M(T_{V1})\right] = 0 \Rightarrow \alpha = \rho\frac{C_Y}{C_X}$$

After replacing value of α in (38), we obtained

$$[M(T_{V1})]_{Min} = \left[\left(L-\frac{1}{n'}\right)-\left(\frac{1}{n}-\frac{2}{n'}+\frac{1}{N}\right)\rho^2\right]S_Y^2$$

Theorem 2:

(5) The estimator T_{V2} in terms of e_1, e_2, e_3 and $e_3^{'}$ is $T_{V2} = \overline{Y}\left[1+e_1+\beta\left\{e_3-e_2+e_1e_3-e_1e_2-\beta e_2e_3+\frac{\beta+1}{2}e_2^2+\frac{\beta-1}{2}e_3^2\right\}\right]$ (39)

(6) The bias of T_{V2} is $B(T_{V2})=\overline{Y}\beta\left(L-\frac{1}{n}\right)\left(\frac{\beta+1}{2}C_X^2-\rho C_Y C_X\right)$ (40)

(7) Mean squared error of T_{V2} is: $M(T_{V2})=\overline{Y}^2\left[\left(L-\frac{1}{n'}\right)C_Y^2+\left(L-\frac{1}{n}\right)\left(\beta_2^2C_X^2-2\beta_2\rho C_Y C_X\right)\right]$ (41)

(8) The minimum m.s.e. of T_{V2} is $[M(T_{V2})]_{Min} = \left[\left(L-\frac{1}{n'}\right)-\left(L-\frac{1}{n}\right)\rho^2\right]S_Y^2$ when $\beta=\rho\frac{C_Y}{C_X}$ (42)

Theorem 3:

(9) The estimator T_{V3} in terms of e_1, e_2, e_3 and $e_3^{'}$ is $T_{V3} = \overline{Y}\left[1+e_1+\gamma\left\{e_3^{'}-e_2-e_1e_2+e_1e_3^{'}-\gamma e_2e_3^{'}+\frac{\gamma+1}{2}e_2^2+\frac{\gamma-1}{2}e_3^{'2}\right\}\right]$ (43)

(10) Bias of T_{V3} is: $B(T_{V3})=\overline{Y}\gamma\left(L-\frac{2}{n'}+\frac{1}{N}\right)\left(\frac{\gamma+1}{2}C_X^2-\rho C_Y C_X\right)$ (44)

(11) Mean squared error of T_{V3} is: $M(T_{V3})=\overline{Y}^2\left[\left(L-\frac{1}{n'}\right)C_Y^2+\left(L-\frac{2}{n'}+\frac{1}{N}\right)\left(\gamma^2C_X^2-2\gamma\rho C_Y C_X\right)\right]$ (45)

(12) The minimum m.s.e. of T_{V3} is: $[M(T_{V3})]_{min} = \left[\left(L-\frac{1}{n'}\right)-\left(L-\frac{2}{n'}+\frac{1}{N}\right)\rho^2\right]S_y^2$, when $\gamma=\rho\frac{C_Y}{C_X}$ (46)

Comparisons

In this section we derived the conditions under which the suggested estimators are superior to existing estimators.

(1) $D_1 = min[M(t_1)] - min[M(T_{V1})] = \left[\frac{1}{n_1}-\frac{1}{N}-L-\frac{1}{n'}\right]S_Y^2 - 2\left[\frac{1}{n^{'}}-\frac{1}{N}\right]\rho^2 S_y^2$ (T_{V1}) is better than t_1,

if $D_1 > 0 \Rightarrow \rho^2 < \frac{1}{2}\frac{\left[\frac{1}{n_1}-\frac{1}{N}-L+\frac{1}{n'}\right]}{\left[\frac{1}{n'}-\frac{1}{N}\right]} \Rightarrow \rho^2 < \frac{1}{2}-\frac{(N-n)(n-n_1)n'N}{nn_1^2(N-1)(N-n')}$, $\Rightarrow \rho < \pm\sqrt{\frac{1}{2}-M} \Rightarrow -\sqrt{\frac{1}{2}-M} < \rho < +\sqrt{\frac{1}{2}-M}$

Where $M < \frac{1}{2} \Rightarrow 2(N-n)(n-n_1)n'N < nn_1^2(N-1)(N-n')$

(2) $D_2 = min[M(t_2)] - min[M(T_{V2})] = \left[\frac{1}{n_1}-\frac{1}{N}-L+\frac{1}{n'}\right]S_Y^2 + \left[L-\frac{1}{r}\right]\rho^2 S_y^2$

(T_{V2}) is better than t_1,

if $D_2 > 0 \Rightarrow \rho^2 < \frac{\left[\frac{1}{n_1}-\frac{1}{N}-L+\frac{1}{n'}\right]}{\left[\frac{1}{n_1}-L\right]} \Rightarrow \rho^2 < \left[\frac{(N-n)(n-n_1)n'N}{nn_1^2(N-1)(N-n')}\right]^{-1}-1 \Rightarrow \rho < \pm\sqrt{\frac{1-M}{M}}$

Where $M < 1 \Rightarrow (N-n)(n-n_1)n'N < nn_1^2(N-1)(N-n')$

(3) $D_3 = \min\left[M\left(t_1\right)\right] - \min\left[M\left(T_{V3}\right)\right] = \left[\frac{1}{n_1} - \frac{1}{N} - L + \frac{1}{n'}\right]\left(1-\rho^2\right)S_Y^2$

$\left(T_{V3}\right)$ is better than t_1, if $D_3 > 0$ $= \left[\frac{1}{n_1} - \frac{1}{N} - L + \frac{1}{n'}\right]\left(1-\rho^2\right)S_Y^2 > 0 \Rightarrow -1 < \rho < 1$

Numerical Illustrations

We consider two populations A and B, first one is the artificial population of size 200 and another one is of size 8306 with the following parameters: [9,10]

Table-1
Parameters of Populations A and B

Population	N	$\overline{Y}$	$\overline{X}$	S_Y^2	S_X^2	ρ	C_X	C_Y
A	200	42.485	18.515	199.0598	48.5375	0.8652	0.3763	0.3321
B	8306	253.75	343.316	338006	862017	0.522231	2.70436	2.29116

Let $n' = 60$, $n = 40$, $n_1 = 35$ for population A and $n' = 2000$, $n = 500$, $n_1 = 450$ for population B respectively. Then the bias and M.S.E of suggested estimators and existing estimators are given in table 2 and 3 for population A and B respectively.

Table-2
Bias and MSE for Population A and B

Estimators	**Population A**		**Population B**	
	Bias	**MSE**	**Bias**	**MSE**
T_{V1}	-0.00181	2.882792	0.256463	478.9972
T_{V2}	0.001983	1.841686	0.050974	561.7505
T_{V3}	0.000174	2.338387	0.307437	458.4694

Table-3
Bias and MSE for Population A and B for Ahmed et al. (2006)

Estimators	**Population A**		**Population B**	
	Bias	**MSE**	**Bias**	**MSE**
$\overline{y}_r$	0	4.692124	0	710.4302
$\overline{y}_{RAT}$	0.00508	4.908211	0.22994	768.7752
$\overline{y}_{COMP}$	0.003879	4.188044	0.050411	689.9429
t_1	0.010856	1.711916	0.43025	537.1631
t_2	0.001939	4.159944	0.050868	689.9452
t_3	0.012795	1.179736	0.481117	516.678

The sampling efficiency of suggested estimators over existing is defined as: $E_i = \frac{\mathrm{Opt}\left[M\left(T_{V1}\right)\right]}{\mathrm{Opt}\left[M\left(t_i\right)\right]}; i = 1,2,3;$ (47)

The efficiency for population A and population B are given in table 4.

Table-4
Efficiency for Population A and B

Efficiency	**Population A**	**Population B**
E_1	1.683957	0.891717
E_2	0.442719	0.814196
E_3	1.982128	0.887341

Discussion

The idea of two-phase sampling is used while considering, the auxiliary population mean is unknown and numbers of available observations are considered as random variable. Some strategies are suggested and the estimators for population mean are derived. Properties of derived estimators like bias and m.s.e are also discussed in this paper. The optimum value of parameters of suggested estimators is obtained as well in same section. Some existing estimators are considered for comparison purpose and two populations A and B considered for numerical study. The sampling efficiency of suggested estimator is calculated and suggested strategy is found very close with existing when $\overline{X}$ is not known.

Conclusions

The proposed estimators are useful when some observations are missing in the sampling and population mean of auxiliary information is unknown. Proposed estimator T_{V2} is found to be more efficient than the existing estimators. The estimators T_{V1} and T_{V3} results are also close with Ahmed estimators.

Acknowledgement

Authors are thankful to the Editorial Board of ISCA-RJRS and referees for their valuable suggestions and for recommending the manuscript for publication.

References

1. Rubin D.B., Inference and missing data, *Biometrica*, **63**, 581-593 **(1976)**
2. Hansen M.H. and Hurwitz W.N., The problem of non-response in Sample Surveys, *Journal of the American Statistical Association*, **41(236)**, 517-529 **(1946)**
3. Sukhatme P.V., Sukhatme B.V., Sukhatme S. and Ashok C., Sampling Theory of Surveys with Applications, *Iowa State University Press, I.S.A.S. Publication*, New Delhi **(1984)**
4. Shukla D., Thakur N. S. and Thakur D. S., Utilization of non-response auxiliary population mean in imputation for missing observations, *Journal of Reliability and Statistical Studies*, **2**, 28-40 **(2009)**
5. Shukla D., Thakur N. S., Thakur D.S. and Pathak S., Linear combination based imputation method for missing data in sample, *International Journal of Modern Engineering Research*, **1(2)**, 580-596 **(2011)**
6. Shukla D., Thakur N. S. and Thakur D. S., Utilization of mixture of $\overline{x}$, $\overline{x}_1$ and $\overline{x}_2$ in imputation for missing data in post-stratification, *African Journal of Mathematics and Computer Science Research*, Vol. 5(4), pp. 78-89, 15 February, **(2012)**
7. Heitjan D.F. and Basu S., Distinguishing 'Missing at random' and 'Missing completely at random', *The American Statistician*, **50**, 207-213 **(1996)**
8. Singh S. and Horn S., Compromised imputation in survey sampling, *Metrika*, **51**, 266-276 **(2000)**
9. Ahmed M.S., Al-Titi O., Al-Rawi Z. and Abu-Dayyeh W., Estimation of a population mean using different imputation methods, *Statistics in Transition*, **7(6)**, 1247-1264 **(2006)**
10. Shukla D. and Thakur N.S., Estimation of mean with imputation of missing data using factor-type estimator, *Statistics in Transition*, **9(1)**, 33-48 **(2008)**
11. Thakur N.S., Yadav K. and Pathak S., Estimation of mean in presence of missing data under two-phase sampling scheme, *Journal of Reliability and Statistical Studies*, **4(2)**, 93-104 **(2011)**
12. Thakur N.S., Yadav K. and Pathak S., Some imputation methods in double sampling scheme for estimation of population mean, *International Journal of Modern Engineering Research*, **2(1)**, 200-207 **(2012)**
13. Thakur N.S., Yadav K. and Pathak S., Mean estimation with imputation in two- phase sampling, *International Journal of Modern Engineering Research,* **2(5),** 3561-3571 **(2012)**
14. Thakur N.S., Yadav K. and Pathak S., Imputation using regression estimators for estimating population mean in two-phase sampling, *Journal of Reliability and Statistical Studies,* **5(2),** 93-104 **(2012)**
15. Shukla D., Thakur N.S. and Pathak S., Some new aspects on imputation in sampling, *African Journal of Mathematics and Computer Science Research,* **6(1)**, 5-15, 15 **(2013)**
16. Shukla D., Thakur N.S., Pathak S. and Rajput D.S., Estimation of mean with imputation of missing data using factor- type estimator in two-phase sampling, *Statistics in Transition*, **10(3)**, 397-414 **(2009)**