On Mean Estimation with Imputation in Two-Phase Sampling Design

Thakur Narendra Singh¹, Yadav Kalpana² and Pathak Sharad³
¹Center for Mathematical Sciences (CMS), Banasthali University, Rajasthan, INDIA
²Department of Statistics, University of Delhi, Delhi, INDIA
³Department of Mathematics and Statistics, Dr. H. S. Gour Central University, Sagar MP, INDIA

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 $\phi$ 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 $\bar{X}$ of auxiliary variable is unknown. A large preliminary simple random sample (without replacement) $S$ of $n$ units is drawn from the...
population \(\Omega = \{1, 2, \ldots, N\}\) to estimate \(\bar{X}\) 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 \(n_i\) responding units and \(n_i = (n-n_i)\) non-responding units. Using the concept of post-stratification, sample may be divided into two groups: responding \((R_i)\) and non-responding \((R_j)\).

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\(^5,6\).

Now it may be consider the population has two types of individuals like \(N_1\) as number of respondents \((R_i)\) and \(N_2\) non-respondents \((R_j)\). 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 \(R_i\) and \(R_j\) groups are expressed as \(W_1 = N_1/N\) and \(W_2 = N_2/N\) such that \(W_1+W_2=1\). Further, let \(Y\) and \(\bar{X}\) be the population means of \(Y\) and \(X\) respectively. For every unit \(i \in R_i\), the value \(y_i\) is observed available. However, for the units \(i \in R_j\), 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 \(i \in R_j\). This is to assume that for sample \(S\), the data \(x_i = \{x_i : i \in S\}\) are known. The following notations are used in the present research manuscript:

\[\bar{x}_i, \bar{y}_i: \text{the sample mean of } X \text{ and } Y \text{ respectively in } S; \bar{x}_i, \bar{y}_i: \text{the sample mean of } X \text{ and } Y \text{ respectively in } R_i; S^i, S^j: \text{ the population mean squares of } X \text{ and } Y \text{ respectively}; C_x, C_y: \text{the coefficient of variation of } X \text{ and } Y\]

\[\rho \text{ Correlation Coefficient in population between } X \text{ and } Y\]

Further, consider few more symbolic representations:

\[L = \left[ 1 - \frac{N-n}{nW_1} \right] \left( 1 - \frac{n}{n} \right) = \left( 1 - \frac{n}{n} \right) \left( 1 - \frac{n}{n} \right)\]

**Large Sample Approximation**

Let \(\bar{y}_i = \bar{Y}(1+e_i); \bar{x}_i = \bar{X}(1+e_i)\) and \(\bar{x} = \bar{X}(1+e_i)\) . which implies the results \(e_i = \frac{\bar{y}_i - \bar{Y}}{\bar{Y}}; e_i = \frac{\bar{x} - \bar{X}}{\bar{X}}\).

\[e_i = \frac{\bar{x} - \bar{X}}{\bar{X}}-1 \text{ and } e_i = \frac{\bar{x} - \bar{X}}{\bar{X}}-1\] Now by using the concept of two-phase sampling and the mechanism of MCAR\(^7\) for given \(n_i, n\) and \(n'\) we have:

\[E(e_i) = E\left( E(e_i) \right) = E\left( \left( \frac{\bar{y}_i - \bar{Y}}{\bar{Y}} \right) \right) = \frac{\bar{y} - \bar{Y}}{\bar{Y}} = 0; \text{ Similarly, } E(e_i) = E(e_i) = E(e_i) = 0\]

\[E(e_i') = E\left( \left( \frac{\bar{y}_i - \bar{Y}}{\bar{Y}} \right) \right) = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x; \text{ Similarly, } E(e_i') = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x; E(e_i') = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x;\]

\[E(e_i) = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x; E(e_i) = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x; E(e_i) = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x;\]

\[E(e_i) = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x; E(e_i) = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x; E(e_i) = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x;\]

\[E(e_i) = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x; E(e_i) = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x; E(e_i) = \left( 1 - \frac{1}{n} \right) \frac{1}{n} C_x;\]
Some Existing Imputation Methods

Let \( \bar{Y} = N^{-1} \sum_{i=1}^{N} y_i \) be the mean of the finite population under consideration. A Simple Random Sampling Without Replacement (SRSWOR), \( S \), of size \( n \) is drawn from \( \{1, 2, \ldots, N\} \) to estimate the population mean \( \bar{Y} \). 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 \( \{y_{i} : i \in S\} \) are known with mean \( \bar{x} = (n)^{-1} \sum_{i=1}^{n} x_i \). Under this setup, some well known imputation methods are given below:

### Mean Method of Imputation:

For \( y_i \) define \( y_i \) as:

\[
\begin{cases} 
    y_i & \text{if } i \in R \\
    \bar{y} & \text{if } i \in R^c
\end{cases}
\]

Using above, the imputation-based estimator of population mean \( \bar{Y} \) is:

\[
\bar{y}_a = \frac{1}{n} \sum_{i \in S} y_i = \bar{y}
\]

The bias and mean squared error is given by

\[
\begin{align*}
\text{(i)} & \quad B(\bar{y}_a) = 0 \\
\text{(ii)} & \quad V(\bar{y}_a) = \left( \frac{1}{r} - \frac{1}{N} \right) S^2_n
\end{align*}
\]

### Ratio Method of Imputation:

For \( y_i \) and \( x_i \), define \( y_i \) as:

\[
\begin{cases} 
    y_i & \text{if } i \in R \\
    \frac{y_i}{\bar{x}} & \text{if } i \in R^c
\end{cases}
\]

where \( \bar{b} = \sum_{i \in S} y_i / \sum_{i \in S} x_i \).

Using above, the imputation-based estimator is:

\[
\bar{y}_r = \frac{1}{n} \sum_{i \in R} y_i = \bar{y}
\]

where \( \bar{y} = \frac{1}{r} \sum_{i \in R} y_i \) and \( \bar{x} = \frac{1}{n} \sum_{i \in S} x_i \).

**Lemma:** The bias and mean squared error of \( \bar{y}_{ras} \) is given by

\[
\begin{align*}
\text{(i)} & \quad B(\bar{y}_{ras}) = P \left( \frac{1}{r} - \frac{1}{n} \right) \left( C^2_x - \rho C_y C_x \right) \\
\text{(ii)} & \quad M(\bar{y}_{ras}) = \left( \frac{1}{n} - \frac{1}{N} \right) S^2_n + \left( \frac{1}{r} - \frac{1}{n} \right) \left[ S^2_n + R^2 + S^2_{y} - 2R, S_{y} \right]
\end{align*}
\]

where \( R = \frac{\bar{Y}}{\bar{X}} \).

### Compromised Method of Imputation:

\[
\begin{cases} 
    (an/r) y_i + (1 - \alpha) \bar{b} \bar{x} & \text{if } i \in R \\
    (1 - \alpha) \bar{b} \bar{x} & \text{if } i \in R^c
\end{cases}
\]
where $\alpha$ is a suitably chosen constant, such that the resultant variance of the estimator is optimum. The imputation-based estimator, for this case, is
\[ \bar{y}_{\text{comp}} = [\alpha \bar{y}_r + (1-\alpha) \bar{y}_x, \bar{x}_r, \bar{x}_x] \] (10)

**Lemma:** The bias, mean squared error and minimum mean squared error at $\alpha = 1 - \rho \frac{C_r}{C_x}$ of $\bar{y}_{\text{comp}}$ is given by

(i) $B(\bar{y}_{\text{comp}}) = \bar{y}(1 - \alpha) \left( \frac{1}{r \cdot n} \right) (C_r - \rho C_x, C_x)$

(ii) $M(\bar{y}_{\text{comp}}) = \left[ \left( \frac{1}{N \cdot n} \right) S_r^2 + \left( \frac{1}{r \cdot n} \right) \left[ S_r^2 + R_r S_r^2 - 2R S_r \right] \right] - \left( \frac{1}{r \cdot n} \right) \alpha \bar{Y} C_r$

(iii) $M(\bar{y}_{\text{comp}})_{\text{min}} = \left[ \left( \frac{1}{r \cdot n} \right) - \left( \frac{1}{r \cdot n} \right) \rho^2 \right] S_r^2$

**Ahmed Methods:**

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

\[ y_n = \left\{ \begin{array}{ll}
    y_i & \text{if } i \in R \\
    \frac{1}{(n-r)} \left[ n y_i \left( \frac{X}{x} \right)^{n-i} - r \bar{y}_r \right] & \text{if } i \in R^c
\end{array} \right. \] (14)

Under this, point estimator is
\[ t_i = y_i \left( \frac{X}{x} \right)^{n-i} \] (15)

**Lemma 1:** The bias, mean squared error and minimum mean squared error at $\beta = \rho \frac{C_r}{C_x}$ of $t_i$ is given by

(i) $B(t_i) = \bar{y} \left( \frac{1}{n} \right) \left( \beta \frac{\beta + 1}{2} C_r - \rho C_x, C_x \right)$

(ii) $M(t_i) = \bar{y} \left( \frac{1}{n} \right) \left( \beta \frac{\beta + 1}{2} C_r + \beta \left( \frac{1}{n} \right) C_r, C_x \right) - \beta \frac{1}{n} \rho C_x, C_x$

(iii) $M(t_i)_{\text{min}} = \left( \frac{1}{n} \right) \left( \beta \frac{\beta + 1}{2} C_r - \beta \rho C_x, C_x \right)$

\[ y_n = \left\{ \begin{array}{ll}
    y_i & \text{if } i \in R \\
    \frac{1}{(n-r)} \left[ n y_i \left( \frac{X}{x} \right)^{n-i} - r \bar{y}_r \right] & \text{if } i \in R^c
\end{array} \right. \] (19)

The point estimator is
\[ t_i = y_i \left( \frac{X}{x} \right)^{n-i} \] (20)

**Lemma:** The bias, mean squared error and minimum mean squared error at $\beta_1 = \rho \frac{C_r}{C_x}$ of $t_i$ is given by

\[ B(t_i) = \left( \frac{1}{r \cdot n} \right) \left( \beta \frac{\beta + 1}{2} \right) C_r - \beta \rho C_x, C_x \] (21)
(ii) \[ M(t_i) = \bar{Y} \left[ \frac{1}{r} - \frac{1}{N} \right] C_i + \beta_1 \left( \frac{1}{r} - \frac{1}{n} \right) C_i - 2\beta_1 \left( \frac{1}{r} - \frac{1}{n} \right) \rho C_i C_s \]  

(22)

(iii) \[ M(t_i) = \left\{ \frac{1}{r} - \frac{1}{N} \right\} S_i^2 - \left( \frac{1}{r} - \frac{1}{n} \right) \frac{S_y}{S_i^2} \]  

(23)

(3) \[ y_{ij} = \begin{cases} y_i & \text{if } i \in R \\ \frac{1}{(n-r)} \left[ n \bar{Y} \left( \bar{X} / x_{ij} \right) \right] & \text{if } i \in R_c \end{cases} \]  

(24)

The point estimator is \[ t_i = \bar{y} \left( \bar{X} / x_{ij} \right) \]  

(25)

**Lemma:** The bias, mean squared error and minimum mean squared error at \( \beta_i = \rho C_i / C_s \) of \( t_i \) is given by

(i) \[ B(t_i) = \left( \frac{1}{r} - \frac{1}{N} \right) \left( \frac{\beta (\beta+1)}{2} \right) C_i - \beta \rho C_i C_s \]  

(26)

(ii) \[ M(t_i) = \left( \frac{1}{r} - \frac{1}{N} \right) \bar{Y} \left[ C_i + \beta C_x - 2\beta \rho C_i C_s \right] \]  

(27)

(iii) \[ M(t_i) = \left( \frac{1}{r} - \frac{1}{N} \right) S_i \left( 1 - \rho \right) \]  

(28)

and when \( k = 4, \beta = 0 \) then \( T_{yw} = t_i = \bar{y}, i \in \{1,2,3\} \)

**Proposed Different Imputation Methods**

Let \( y_{ij} \) denotes the \( i \)th available observation for the \( j \)th imputation. We suggest the following imputation methods:

(4) \[ y_{ij} = \begin{cases} y_i & \text{if } i \in R_i \\ \bar{y} - \frac{1}{1-W_i} \left[ \frac{\bar{X}}{x_{ij}} \right]^a - W_i & \text{if } i \in R_c \end{cases} \]  

(29)

where \( \alpha \) is suitably chosen constant, such that the variance the resultant estimator is minimum. Under this method, the point estimator of \( \bar{Y} \) is \[ T_{iv} = \bar{y} \left( \frac{\bar{X}}{x_{ij}} \right)^a \]  

(30)

(5) \[ y_{ij} = \begin{cases} y_i & \text{if } i \in R_i \\ \bar{y} - \frac{1}{1-W_i} \left[ \frac{\bar{x}_{ij}}{x_{ij}} \right] - W_i & \text{if } i \in R_c \end{cases} \]  

(31)

where \( \beta \) is suitably chosen constant, such that the variance the resultant estimator is minimum. Under this method, the point estimator of \( \bar{Y} \) is \[ T_{iv} = \bar{y} \left( \frac{\bar{x}_{ij}}{x_{ij}} \right)^a \]  

(32)
(6) \[ y_{ri} = \begin{cases} \frac{y_i}{(1-W_i)} \left( \frac{x_i}{x_i} \right)^\gamma - W_i & \text{if } i \in R_1 \\ \left( \frac{x_i}{x_i} \right)^\gamma - W_i & \text{if } i \in R_2 \end{cases} \] (33)

Where \( \gamma \) is suitably chosen constant, such that the variance of the resultant estimator is minimum. Under this method, the point estimator of \( \bar{Y} \) is \( T_{ri} = y_i \left( \frac{x_i}{x_i} \right)^\gamma \) (34)

**Note:** At \( \gamma = 1 \) (−1) then the estimator \( T_{ri} \) 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 \( T_{ri} \) in terms of \( e_i^* \), \( i = 1, 2, 3 \) and \( e_i \), could be expressed up to first order of approximation:
   \[ T_{ri} = \bar{Y} \left[ 1 + e_i + \alpha \left( e_i - e_i^* - e_i e_i^* + e_i e_i^* - \alpha e_i e_i^* + \frac{\alpha + 1}{2} e_i^3 + \frac{\alpha - 1}{2} e_i^3 \right) \right] \] (35)

   **Proof:**
   \[ T_{ri} = \bar{Y} \left( 1 + e_i + \frac{\alpha (\alpha - 1)}{2} e_i^2 \right) \left( 1 - \alpha e_i + \frac{\alpha (\alpha + 1)}{2} e_i^2 \right) \]
   \[ = \bar{Y} \left[ 1 + e_i + \alpha \left( e_i - e_i^* - e_i e_i^* + e_i e_i^* - \alpha e_i e_i^* + \frac{\alpha + 1}{2} e_i^3 + \frac{\alpha - 1}{2} e_i^3 \right) \right] \]

2. Bias of \( T_{ri} \) is: \( B(T_{ri}) = \bar{Y} \alpha \left( 1 - \frac{2}{n} + \frac{1}{N} \right) \left( \frac{\alpha + 1}{2} C_i^2 - \rho C_x C_i \right) \) (36)

   **Proof:**
   \[ B(T_{ri}) = E[T_{ri} - \bar{Y}] = \bar{Y} E \left[ 1 + e_i + \alpha \left( e_i - e_i^* - e_i e_i^* + e_i e_i^* - \alpha e_i e_i^* + \frac{\alpha + 1}{2} e_i^3 + \frac{\alpha - 1}{2} e_i^3 \right) \right] \]
   \[ = \bar{Y} \alpha \left( 1 - \frac{2}{n} + \frac{1}{N} \right) \left( \frac{\alpha + 1}{2} C_i^2 - \rho C_x C_i \right) \]

3. Mean squared error of \( T_{ri} \), up to first order of approximation could be written as:
   \[ M(T_{ri}) = \bar{Y} \left[ \left( L - \frac{1}{n} \right) C_i^2 + \left( \frac{1}{n} - \frac{2}{n} + \frac{1}{N} \right) \left( \alpha C_x^2 - 2 \rho C_x C_i \right) \right] \] (37)

   **Proof:**
   \[ M(T_{ri}) = E \left[ T_{ri}^2 - \bar{Y}^2 \right] = \bar{Y} \left[ 1 + e_i + \alpha \left( e_i - e_i^* - e_i e_i^* + e_i e_i^* - \alpha e_i e_i^* + \frac{\alpha + 1}{2} e_i^3 + \frac{\alpha - 1}{2} e_i^3 \right) \right]^2 \]
   \[ = \bar{Y} \left[ \left( L - \frac{1}{n} \right) C_i^2 + \left( \frac{1}{n} - \frac{2}{n} + \frac{1}{N} \right) \left( \alpha C_x^2 - 2 \rho C_x C_i \right) \right] \]

4. Minimum mean squared error of \( T_{ri} \) is:
   \[ [M(T_{ri})]_{\alpha} = \left( L - \frac{1}{n} \right) \left( \frac{1}{n} - \frac{2}{n} + \frac{1}{N} \right) \rho^2 S_i^2 \text{ when } \alpha = \rho \left( \frac{C_i}{C_x} \right) \] (38)

   **Proof:** First differentiate (38) with respect to \( \alpha \) and then equate to zero, we get
   \[ \frac{d}{d \alpha} [M(T_{ri})] = 0 \Rightarrow \alpha = \rho \left( \frac{C_i}{C_x} \right) \]
After replacing value of $\alpha$ in (38), we obtained
\[ [M(T_{v1})]_{\alpha} = \left[ \left\{ 1 - \frac{1}{N} \right\} - \frac{1}{n} \frac{2}{N} + \frac{1}{N} \right] \rho^2 \] (39)

**Theorem 2:**

(5) The estimator $T_{v1}$ in terms of $e_1, e_2, e_3$, and $e_4$ is:
\[ \hat{T}_{v1} = Y \left[ 1 + e_1 + \beta \left( e_1 - e_2 + e_4 - e_3 - \beta e_2 e_3 \right) \right] \]
(39)

(6) The bias of $T_{v1}$ is:
\[ B(T_{v1}) = \frac{1}{n} \left( \beta \cdot C^2 - \rho C \cdot C_4 \right) \]
(40)

(7) Mean squared error of $T_{v1}$ is:
\[ M(T_{v1}) = \left[ \left\{ 1 - \frac{1}{n} \right\} \rho^2 \right] \]
(41)

(8) The minimum m.s.e. of $T_{v1}$ is:
\[ \left[ M(T_{v1}) \right]_{\alpha} = \left[ \left\{ 1 - \frac{1}{n} \right\} \rho^2 \right] S^2_\alpha \quad \text{when} \quad \beta = \frac{C_4}{C_4} \]
(42)

**Theorem 3:**

(9) The estimator $T_{v2}$ in terms of $e_1, e_2, e_3$, and $e_4$ is:
\[ \hat{T}_{v2} = Y \left[ 1 + e_1 + \rho \left( e_1 - e_2 + e_4 - e_3 \right) \right] \]
(43)

(10) Bias of $T_{v2}$ is:
\[ B(T_{v2}) = \frac{1}{n} \left( \rho \cdot C^2 - \rho \cdot C_4 \cdot C_4 \right) \]
(44)

(11) Mean squared error of $T_{v2}$ is:
\[ M(T_{v2}) = \left[ \left\{ 1 - \frac{1}{n} \right\} \rho^2 \right] \]
(45)

(12) The minimum m.s.e. of $T_{v2}$ is:
\[ \left[ M(T_{v2}) \right]_{\alpha} = \left[ \left\{ 1 - \frac{1}{n} \right\} \rho^2 \right] S^2_\alpha \quad \text{when} \quad \gamma = \frac{C_4}{C_4} \]
(46)

**Comparisons**

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

(1) $D_1 = \min [M(t_1)] - \min [M(T_{v1})] = \left[ \left\{ 1 - \frac{1}{n} \right\} \frac{1}{n} - L \frac{1}{n} \right] S^2_\alpha - \left[ \left\{ 1 - \frac{1}{n} \right\} \frac{1}{n} \right] \rho^2 \ S^2_\alpha$; $T_{v1}$ is better than $t_1$, if $D_1 > 0 \Rightarrow \rho^2 < \frac{1}{2} \left( \frac{1 - \frac{1}{n} - L + \frac{1}{n}}{1 - \frac{1}{n}} \right)$

(2) $D_2 = \min [M(t_1)] - \min [M(T_{v2})] = \left[ \left\{ 1 - \frac{1}{n} \right\} \frac{1}{n} - L + \frac{1}{n} \right] S^2_\alpha + \left[ \left\{ 1 - \frac{1}{n} \right\} \frac{1}{n} \right] \rho^2 \ S^2_\alpha$

$T_{v2}$ is better than $t_1$, if $D_2 > 0 \Rightarrow \rho^2 < \frac{1}{2} \left( \frac{1 - \frac{1}{n} - L + \frac{1}{n}}{1 - \frac{1}{n} - L} \right)$

Where $M < \frac{1}{2} \Rightarrow 2(N - n)(n - n_1)nN < m_{n_1}(N - 1)(N - n)$
\((3)\) \[ D_3 = \min[M(t_1)] - \min[M(T_{V3})] = \left[ \frac{1}{n_1} - \frac{1}{N} - L + \frac{1}{n} \right] \left( 1 - \rho^2 \right) s^2_Y \]

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

**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:

<table>
<thead>
<tr>
<th>Parameters of Populations A and B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
</tbody>
</table>

Let \(n = 60\), \(n = 40\), \(n = 35\) for population A and \(n = 2000\), \(n = 500\), \(n = 450\) 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>
<thead>
<tr>
<th>Estimators</th>
<th>Population A</th>
<th>Population B</th>
</tr>
</thead>
<tbody>
<tr>
<td>T_{v1}</td>
<td>-0.000181</td>
<td>2.882792</td>
</tr>
<tr>
<td>T_{v2}</td>
<td>0.001983</td>
<td>1.841686</td>
</tr>
<tr>
<td>T_{v3}</td>
<td>0.000174</td>
<td>2.333837</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimators</th>
<th>Population A</th>
<th>Population B</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\bar{y})</td>
<td>0</td>
<td>4.692124</td>
</tr>
<tr>
<td>(\bar{y}_{est})</td>
<td>0.00508</td>
<td>4.908211</td>
</tr>
<tr>
<td>(\bar{y}_{cone})</td>
<td>0.003879</td>
<td>4.188044</td>
</tr>
<tr>
<td>(t_{i})</td>
<td>0.010856</td>
<td>1.711916</td>
</tr>
<tr>
<td>(t_{2})</td>
<td>0.001939</td>
<td>4.159944</td>
</tr>
<tr>
<td>(t_{3})</td>
<td>0.012795</td>
<td>1.179736</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Efficiency</th>
<th>Population A</th>
<th>Population B</th>
</tr>
</thead>
<tbody>
<tr>
<td>(E_i)</td>
<td>1.683957</td>
<td>0.891717</td>
</tr>
<tr>
<td>(E_i)</td>
<td>0.442719</td>
<td>0.814196</td>
</tr>
<tr>
<td>(E_i)</td>
<td>1.982128</td>
<td>0.887341</td>
</tr>
</tbody>
</table>
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 $\bar{X}$ 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 $T_{V_2}$ is found to be more efficient than the existing estimators. The estimators $T_{V_1}$ and $T_{V_3}$ 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