ESTIMATION OF MISSING OBSERVATIONS USING AUXILIARY INFORMATION AND POPULATION MEAN UNDER ADAPTIVE CLUSTER SAMPLING

Shubhangi Chaurasia, R. K. Shrivastava

1. SMS Govt. Model Science College, Gwalior, M.P.
2. Principal, Dr. Bhagwat Sahay Govt.College, Gwalior, M.P.

Abstract

In this paper, we have proposed some estimation strategies to estimate missing data in the sample under adaptive sampling scheme and then obtained point estimator for population mean. All the proposed estimators are found to be biased and therefore, the expression of bias, mean squared error and optimum mean squared error are derived in terms of population parameters up to first order of approximation. For testing the performance of the proposed estimators, an empirical study is performed over a dataset as well.

Keywords: Imputation, Adaptive Cluster Sampling (ACS), Ratio type estimator, Bias, Mean squared error (MSE).

1. Introduction

If in a survey, the sampling units are rare and hidden clustered in the population then designing an efficient strategy is challenging and usual strategies may be less efficient. In such cases, adaptive cluster sampling is more recommendable than other existing methodologies. "In adaptive cluster sampling, if the i^{th} unit of the sample S of size n satisfied a pre-specified condition $\mathcal{A}$, then the neighboring units of the population are also included in the sample. Further, if the neighbors of neighboring units are also satisfying the condition $\mathcal{A}$ then these neighbors of neighboring units are also added to the sample. This process is continued until all the neighboring units of the population satisfying condition $\mathcal{A}$ in respect to i^{th} selected unit are selected in the sample" **Borkowski and Turk (2015) [1]**. The set of every i^{th} unit of the sample satisfying the condition $\mathcal{A}$ constitutes a network of size m_i, i.e. $\mathcal{N}_i = \{i: i \in \mathcal{A}\}$. Obviously, the size of the sample S varies due to adding neighboring units in the sample that breaks the assumption of prefix sample size. For dealing such drawback the mean of the variables of the i^{th} network is obtained and considered i^{th} sample observation as discussed in **Chutiman (2013)[2]**. For a more through review of ACS, see **Turk and Borkowski (2005) [19].**

Let a finite population $U = \{U_1, U_2, U_3, \dots, U_N\}$ of size N is under consideration for assessment. Let the variable x is the study variable and variable y be the auxiliary variable for this case. Let a preliminary sample S of size n is drawn from this population and then unit $i \in S$ construct the network $\mathcal{N}_i$ of size m_i for each unit $i \in S$.

$$\bar{x}_{i\cdot} = \frac{1}{m_i}\sum_{j=1}^{m_i} x_{ij}\ ; \qquad \bar{y}_{i\cdot} = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}\,.$$

Then the final sample of size n is considered as $(\bar{x}_{i\cdot}, \bar{y}_{i\cdot})$; $i = 1, 2, 3, \dots, n$ and the estimate of population mean under adaptive cluster sampling is given as

$(\bar{x}_{..}^{n}, \bar{y}_{..}^{n})$, where, $\bar{x}_{..}^{n} = \frac{1}{n}\sum_{i=1}^{n} \bar{x}_{i.}$, and $\bar{y}_{..}^{n} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i.}$ respectively. Since the sample S of size n is simple random sample, therefore, $E(\bar{x}_{..}) = \bar{X}$ and $E(\bar{y}_{..}) = \bar{Y}$, where, $\bar{X}$ and $\bar{Y}$ are the population mean and defined as $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ respectively. Also, the variance of $(\bar{x}_{..}^{n}, \bar{y}_{..}^{n})$ are defined as $\bar{x}_{..}^{n} = \left(\frac{1}{n} - \frac{1}{N}\right) S_x^2$ and $\bar{y}_{..}^{n} = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$ respectively. The symbols S_x^2 and S_y^2 are defined as $S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ and $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ respectively.

Table 1.1: Layout of the sample S and networks N_i in adaptive cluster sampling

	Selected units in the final sample S of size n								
	1	**2**	**3**	...	...	i	...	...	n
Network N_i of size m_i	u_{11}	u_{21}	u_{31}	...	...	u_{i1}	...	...	u_{n1}
	u_{12}	u_{22}	u_{32}	...	...	u_{i2}	...	...	u_{n2}
	u_{13}	u_{23}	u_{33}	...	...	u_{i3}	...	...	u_{n3}
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	u_{1j}	u_{2j}	u_{3j}	...	...	u_{ij}	...	...	u_{nj}
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	u_{1m_1}	u_{2m_2}	u_{3m_3}	...	...	u_{im_i}	...	...	u_{nm_n}

Table 1.2: Corresponding observations of sample S and networks N_i in adaptive sampling

	Observations of the final sample S of size n								
	1	2	3	...	...	i	...	...	n
Observations in Network N_i of size m_i	(x_{11}, y_{11})	(x_{21}, y_{21})	(x_{31}, y_{31})	...	...	(x_{i1}, y_{i1})	...	...	(x_{n1}, y_{n1})
	(x_{12}, y_{12})	(x_{22}, y_{22})	(x_{32}, y_{32})	...	...	(x_{i2}, y_{i2})	...	...	(x_{n2}, y_{n2})
	(x_{13}, y_{13})	(x_{23}, y_{23})	(x_{33}, y_{33})	...	...	(x_{i3}, y_{i3})	...	...	(x_{n3}, y_{n3})
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	(x_{1j}, y_{1j})	(x_{2j}, y_{2j})	(x_{3j}, y_{3j})	...	...	(x_{ij}, y_{ij})	...	...	(x_{nj}, y_{nj})
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	(x_{1m_1}, y_{1m_1})	(x_{2m_2}, y_{2m_2})	(x_{3m_3}, y_{3m_3})	...	...	(x_{im_i}, y_{im_i})	...	...	(x_{nm_n}, y_{nm_n})
Network's Mean	$(\bar{x}_{1.}, \bar{y}_{1.})$	$(\bar{x}_{2.}, \bar{y}_{2.})$	$(\bar{x}_{3.}, \bar{y}_{3.})$	...	...	$(\bar{x}_{i.}, \bar{y}_{i.})$	...	...	$(\bar{x}_{n.}, \bar{y}_{n.})$

If some observations of study variable in the sample S are missing then the estimation of $\bar{X}$ is not possible due to missingness of the observations in the sample and we have a question - how to estimate the missing observations and then $\bar{X}$? For solving the issue a procedure called imputation is applied, which uses the available data as a tool for the replacement of the missing values as discussed in - Missing data estimation based on the chaining technique in survey sampling **(Thakur and Shukla (2022) [15].** A detailed discussion on missing observation see **Shukla and Thakur (2008) [12], Thakur et. al. (2011)[16], Thakur et.al (2014) [17] and Thakur et.al (2016) [18].**

Let in an adaptive sample of size n, there are only r observations available in the sample and $(n-r)$ observations are missing $(r<n)$. Let the symbol R be the set of available observations and the complementary set of R (viz., R^C) represents the missing observations such that $S = R \cup R^C$. Let for every observation $i \epsilon R$, the value $\bar{x}_{i.}$ is available and for the observations $i \epsilon R^C$ the observations $\bar{x}_{i.}$ are missing and we need to assess these missing values, either by guessing or by statistical techniques. Imputation is such a technique which used for estimating missing data. Also, let the observations of auxiliary variable $\bar{y}_{i.}$ are available completely in the sample, i.e. the data $y_s = \{\bar{y}_{i.} : i \epsilon S\}$ are known. Under this setup, we cover the following imputation methods under adaptive cluster sampling:

(1) Mean Method of Imputation: For sample values $\bar{x}_{i.}$ and $\bar{y}_{i.}$, define the j^{th} imputation method for only study variable as

$$(\bar{x}_{i.})_1 = \begin{cases} \bar{x}_{i.} & if \quad i \epsilon R \\ \bar{x}^r_{..} & if \quad i \epsilon R^C \end{cases} \tag{1.1}$$

The imputation-based estimator of population mean under adaptive cluster sampling is

$$T_1 = \bar{x}^r_{..} \tag{1.2}$$

Hence, the mean of the available data is representing the estimator of population mean.

(2) Ratio Method of Imputation: For sample values $\bar{x}_{i.}$ and $\bar{y}_{i.}$, define the j^{th} imputation method as

$$(\bar{x}_{i.})_2 = \begin{cases} \bar{x}_{i.} & if \quad i \epsilon R \\ \hat{b}\bar{y}_{i.} & if \quad i \epsilon R^C \end{cases} \tag{1.3}$$

where, $\hat{b} = \frac{\sum_{i \epsilon R} \bar{x}_{i.}}{\sum_{i \epsilon R} \bar{y}_{i.}}$.

Under this setup the imputation-based estimator of population mean under adaptive cluster sampling is

$$T_2 = \frac{\bar{x}^r_{..}}{\bar{y}^r_{..}} \bar{y}^n_{..} \tag{1.4}$$

(3) Ahmed Methods of Imputation:

(A) For sample values $\bar{x}_{i.}$ and $\bar{y}_{i.}$, define the j^{th} imputation method as

$$(\bar{x}_{i.})_3 = \begin{cases} \bar{x}_{i.} & if \quad i \in R \\ \frac{1}{n-r}\left[n\bar{x}_{..}^{r}\left(\frac{\bar{Y}}{\bar{y}_{..}^{n}}\right)^{p} - r\bar{x}_{..}^{r}\right] & if \quad i \in R^{C} \end{cases} \tag{1.5}$$

where, p is a suitably chosen constant so that the mean squared error of the resultant estimator is optimum. Under this imputation strategy the point estimator of population mean is defined as

$$T_3 = \bar{x}_{..}^{r}\left(\frac{\bar{Y}}{\bar{y}_{..}^{n}}\right)^{p} \quad \ldots \quad (1.6)$$

(B) For sample values $\bar{x}_{i.}$ and $\bar{y}_{i.}$, define the j^{th} imputation method as

$$(\bar{x}_{i.})_4 = \begin{cases} \bar{x}_{i.} & if \quad i \in R \\ \frac{1}{n-r}\left[n\bar{x}_{..}^{r}\left(\frac{\bar{y}_{..}^{n}}{\bar{y}_{..}^{r}}\right)^{q} - r\bar{x}_{..}^{r}\right] & if \quad i \in R^{C} \end{cases} \quad \ldots \quad (1.7)$$

where, q is a suitably chosen constant so that the mean squared error of the resultant estimator is optimum. Under this imputation strategy the point estimator of population mean is defined as

$$T_4 = \bar{x}_{..}^{r}\left(\frac{\bar{y}_{..}^{n}}{\bar{y}_{..}^{r}}\right)^{q} \quad \ldots \quad (1.8)$$

(C) For sample values $\bar{x}_{i.}$ and $\bar{y}_{i.}$, define the j^{th} imputation method as

$$(\bar{x}_{i.})_5 = \begin{cases} \bar{x}_{i.} & if \quad i \in R \\ \frac{1}{n-r}\left[n\bar{x}_{..}^{r}\left(\frac{\bar{Y}}{\bar{y}_{..}^{r}}\right)^{s} - r\bar{x}_{..}^{r}\right] & if \quad i \in R^{C} \end{cases} \quad \ldots \quad (1.9)$$

where, s is a suitably chosen constant so that the mean squared error of the resultant estimator is optimum. Under this imputation strategy the point estimator of population mean is defined as

$$T_5 = \bar{x}_{..}^{r}\left(\frac{\bar{Y}}{\bar{y}_{..}^{r}}\right)^{s} \quad \ldots \quad (1.10)$$

2. Review of Literature

In the literature of sampling theory, a number of estimation methods are available under adaptive cluster sampling design. Some of the commonly used estimation methods either considering only study variable x or using the constants of supplementary or auxiliary variable y are, unit mean, ratio method, product method, regression method etc. as given by **Chutiman (2013)[2].** "Ratio estimator in ACS without replacement of network is more

efficient than the ratio estimators in ACS without replacement of units." **Chutiman and Chiangpradit (2014) [3]**. Improved ratio estimators of population mean in adaptive cluster sampling is suggested by **Yadav et.al. (2016) [20].**

According to availability of auxiliary variable y or other circumstances the adaptive cluster sampling is classified into several adaptive sampling designs. Different type of adaptive cluster sampling designs, like - adaptive cluster double sampling design, negative adaptive cluster sampling design, two-stage negative adaptive cluster sampling design, etc. are reviewed and explained very well by **Medina and Thompson (2004) [10]**, **Latpate and Kshirsagar (2018a) [5]** and **Latpate and Kshirsagar (2018b) [6]** respectively. A brief review of ACS is performed by **Narkhede et.al. (2019) [11].**

An efficient sampling technique known as adoptive cluster sampling useful for rare and hidden clustered populations, like – animals or plants of rare and endangered species, fisheries, uneven minerals exploration, pollutions concentrations, epidemiology of sporadic diseases, noise problems, drug users, HIV, AIDS etc. patients, criminals and hot spot investigations, etc. proposed by **Singh and Yadav (2019)[14]**. Another example is explained in **Gattone et.al. (2016) [4]** for negatively correlated data where wildlife population in a protected area can partly be influenced by such factors as disease and pollution where the presence of wildlife diseases or higher environmental pollution decreases population totals and the distribution of wildlife. **Smith et.al. (2003) [13]** applied ACS in a survey of freshwater mussels which tend to be spatially clustered.

Some other more reviews on ACS can also be also seen in **Latpate and Kshirsagar (2018c) [7], (2018d) [8] and (2018e) [9]**.

3. Large sample approximations

Let $\bar{x}_{..}^{r} = \bar{X}(1+\varepsilon)$, $\bar{y}_{..}^{r} = \bar{Y}(1+\delta)$ and $\bar{y}_{..}^{n} = \bar{Y}(1+\eta)$ for large r and n. Using the concept of two-phase sampling and missing completely at random for given r and n, we have

$E(\varepsilon) = 0,$ $E(\delta) = 0,$ $E(\eta) = 0$

$E(\varepsilon^2) = \left(\frac{1}{r} - \frac{1}{N}\right) C_X^2$, $E(\delta^2) = \left(\frac{1}{r} - \frac{1}{N}\right) C_Y^2,$ $E(\eta^2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_Y^2$

$E(\varepsilon\delta) = \left(\frac{1}{r} - \frac{1}{N}\right) \rho_{XY} C_X C_Y$, $E(\varepsilon\eta) = \left(\frac{1}{n} - \frac{1}{N}\right) \rho_{XY} C_X C_Y$, $E(\delta\eta) = \left(\frac{1}{n} - \frac{1}{N}\right) C_Y^2$

Also, $E\left(\propto^i \beta^j\right) = 0, \quad \forall\, i+j > 2,$ where $\propto, \beta = \varepsilon, \delta$ or η and i, j are integers.

where $C_X^2 = \frac{S_X^2}{\bar{X}^2}$, $C_y^2 = \frac{S_{XY}^2}{\bar{Y}^2}$, $\rho = \frac{S_{xy}}{S_x S_y}$ and S_X^2, S_X^2 and S_{xy} have their usual meanings.

$M_1 = \left(\frac{1}{r} - \frac{1}{N}\right)$, $M_2 = \left(\frac{1}{n} - \frac{1}{N}\right)$ and $M_3 = (M_1 - M_2) = \left(\frac{1}{r} - \frac{1}{n}\right)$.

4. Characteristics of proposed estimators

Denote the symbols $B(.)$ and $M(.)$ for bias and mean squared error of the estimator under consideration respectively. Using the concept of large sample approximations and

mathematical expectations described in previous section we derived the bias and mean squared errors of proposed estimators as given in the form of the following theorems.

Theorem 4.1: The estimator T_1 is unbiased and its variance is given by

$$Var(T_1) = \frac{1}{\bar{X}^2}\left(\frac{1}{r} - \frac{1}{N}\right) S_X^2$$

$$= \left(\frac{1}{r} - \frac{1}{N}\right) C_X^2 \qquad (4.1)$$

Theorem 4.2:

i. The estimator T_2 can be expressed in terms of population parameter up to first order approximation is

$$T_2 = \bar{X}\,[1 + \epsilon + \eta + \epsilon\eta - \delta - \epsilon\delta - \delta\eta - \epsilon\eta\delta + \delta^2 + \epsilon\delta^2] \qquad (4.2)$$

ii. The bias of estimator T_2 is given by

$$B(T_2) = -\,[M_2 C_Y^2 + M_3 \rho_{XY} C_X C_Y\,] \qquad (4.3)$$

iii. The mean square error of T_2 is given by

$$M(T_2) = \bar{X}^2[\,M_1 C_X^2 + \; M_3(C_Y^2 - 2\,\rho_{XY} C_X C_Y)] \qquad (4.4)$$

PROOFS:

i. $T_2 = \frac{\bar{x}_{..}^{r}}{\bar{y}_{..}^{r}}\bar{y}_{..}^{n} = \frac{\bar{X}(1+\epsilon)}{\bar{Y}(1+\delta)}\bar{Y}(1+\eta)$ [by using first order approximation]

$$= \bar{X}(1+\epsilon)(1+\eta)(1+\delta)^{-1}$$
$$= \bar{X}(1+\epsilon)(1+\eta)(1-\delta+\delta^2-\delta^3 \ldots\ldots\ldots)$$
$$= \bar{X}(1+\epsilon+\eta+\epsilon\eta-\delta-\epsilon\delta-\delta\eta-\epsilon\eta\delta+\delta^2+\epsilon\delta^2)$$

ii. The bias of the estimator T_2 in terms of population parameters upto first order can be expressed as:

$$B(T_2) = E[T_2 - \bar{X}]$$
$$= E[\bar{X}(1+\epsilon+\eta+\epsilon\eta-\delta-\epsilon\delta-\delta\eta-\epsilon\delta\eta+\cdots) - \bar{X}]$$
$$= E[\epsilon+\eta+\epsilon\eta-\delta-\epsilon\delta-\delta\eta-\epsilon\delta\eta+\cdots]$$

By ignoring higher terms of order (n^{-1}) and taking expectations,

$$B(T_2) = E(\epsilon\eta) - E(\epsilon\delta) - E(\delta\eta)$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{XY} C_X C_Y - \left(\frac{1}{r} - \frac{1}{N}\right)\rho_{XY} C_X C_Y - \left(\frac{1}{n} - \frac{1}{N}\right) C_y^2$$
$$= -\left(\frac{1}{n} - \frac{1}{N}\right) C_Y^2 - \left(\frac{1}{r} - \frac{1}{n}\right)\rho_{XY} C_X C_Y$$

$$\therefore B(T_2) = -\,[M_2 C_Y^2 + M_3 \rho_{XY} C_X C_Y\,]$$

iii. The mean squared error of the estimator T_2 in terms of population parameters upto first order can be written as:

$$MSE(T_2) = E\,[T_2 - \bar{X}]^2$$

$$= E\,[\bar{X}(1 + \epsilon + \eta + \epsilon\eta - \delta - \epsilon\delta - \delta\eta - \epsilon\delta\eta + \cdots \ldots \ldots .) - \mu_Y]^2$$
$$= E\,[\bar{X}^2(\epsilon + \eta + \epsilon\eta - \delta - \epsilon\delta - \delta\eta - \epsilon\delta\eta + \cdots \ldots \ldots \ldots \ldots)^2]$$
$$= \bar{X}^2 E\,[\epsilon + \eta + \epsilon + \eta - \delta - \epsilon\delta - \delta\eta - \epsilon\delta\eta + \cdots \ldots \ldots \ldots .]^2$$

By squaring and taking expectations and ignoring higher terms of $O(n^{-1})$

$$MSE(T_2) = \bar{X}^2[E(\epsilon^2) + E(\eta^2) + E(\delta^2) + 2E(\epsilon\eta) - 2E(\epsilon\delta) - 2E(\delta\eta)]$$

$$= \bar{X}^2\left[\left(\frac{1}{r} - \frac{1}{N}\right)C_X^2 - \left(\frac{1}{n} - \frac{1}{N}\right)C_Y^2 + \left(\frac{1}{r} - \frac{1}{N}\right)C_Y^2 + 2\left(\frac{1}{n} - \frac{1}{N} - \frac{1}{r} + \frac{1}{N}\right)\rho_{XY}C_XC_Y\right]$$

$$= \bar{X}^2\left[\left(\frac{1}{r} - \frac{1}{N}\right)C_X^2 + \left(\frac{1}{r} - \frac{1}{n}\right)(C_Y^2 - 2\rho_{XY}C_XC_Y)\right]$$

$$= \bar{X}^2[\,M_1C_X^2 + M_3(C_Y^2 - 2\,\rho_{XY}C_XC_Y)]$$

Theorem 4.3:

i. The estimator T_3 in the relation of ε, η and δ is given by up to first order approximation is

$$T_3 = \bar{X}\left[1 + \varepsilon - p\eta - p\varepsilon\eta + \frac{p\,(p+1)}{2}\,\eta^2\right] \quad (4.5)$$

ii. The bias of estimator T_3 is given by

$$B(T_3) = M_2\bar{X}\left[\frac{p\,(p+1)}{2}C_Y^2 - p\rho_{XY}C_XC_Y\right] \quad (4.6)$$

iii. The mean square error of T_3 is given by

$$M(T_3) = \bar{X}^2[\,M_1C_X^2 + p^2\,M_2C_Y^2 - 2p\,M_2\rho_{XY}C_XC_Y] \quad (4.7)$$

iv. The minimum mean squared error of T_3 is given by

$$M(T_3)_{min} = M_1S_X^2 - M_2\rho_{XY}^2\,S_X^2 \quad (4.8)$$

for the optimum value of $p = \rho_{XY}\frac{C_X}{C_Y}$

PROOF:

i. $T_3 = \bar{x}_{..}^{r}\left(\frac{\bar{Y}}{\bar{y}_{..}^{n}}\right)^p$

$= \bar{X}\,(1 + \varepsilon)\left(\frac{\bar{Y}}{\bar{Y}\,(1+\eta)}\right)^p$ [by using first order approximation]

$= \bar{X}\,(1 + \varepsilon)(1 + \eta)^{-p}$

$= \bar{X}\,(1 + \varepsilon)\left[1 - p\eta + \frac{p\,(p+1)}{2}\,\eta^2 - \cdots\right]$

$= \bar{X}\left[1 + \varepsilon - p\eta - p\varepsilon\eta + \frac{p\,(p+1)}{2}\,\eta^2\right]$[ignoring higher terms of order n^{-1}]

ii. $B(T_3) = E(T_3 - \bar{X}) = \bar{X}\,E\left(\varepsilon - p\eta - p\varepsilon\eta + \frac{p\,(p+1)}{2}\,\eta^2\right)$

$= \bar{X}\left[-p\,E(\varepsilon\eta) + \frac{p\,(p+1)}{2}\,E(\eta^2)\right]$

$$= \bar{X}\left[-p\, M_2\rho_{XY}C_XC_Y + \frac{p\,(p+1)}{2}\, M_2C_Y^2\right]$$

$$= M_2\bar{X}\left[\frac{p\,(p+1)}{2}C_Y^2 - p\rho_{XY}C_XC_Y\right]$$

iii. $M(T_3) = E(T_3 - \bar{X})^2 = \bar{X}^2E[\varepsilon - p\eta]^2$

$= \bar{X}^2\,[E(\varepsilon)^2 - 2p\,E(\varepsilon\eta) + p^2E(\eta^2)]$

$= \bar{X}^2[M_1C_X^2 + p^2M_2C_Y^2 - 2pM_2\rho_{XY}C_XC_Y]$

iv. Minimum m.s.e. occurs when

$$\frac{d}{dp}M(T_3) = 0$$

$$\frac{d}{dp}\{\bar{X}^2[M_1C_X^2 + p^2M_2C_Y^2 - 2pM_2\rho_{XY}C_XC_Y]\} = 0$$

we get optimum value of p which is given by

$$p = \rho_{XY}\frac{C_X}{C_Y}$$

on putting this value of p in equation (4.6), we get

$$M(T_3)_{min} = M_1S_X^2 - M_2\rho_{XY}^2\,S_X^2$$

Theorem 4.4:

i. The estimator T_4 in the relation of ε, η and δ is given by up to first order approximation is

$$T_4 = \bar{X}\left[1 + \varepsilon + q\eta - q\delta + q\varepsilon\eta - q\varepsilon\delta - q^2\delta\eta + \frac{q\,(q+1)}{2}\,\eta^2 + \frac{q\,(q+1)}{2}\,\delta^2\right] \quad (4.9)$$

ii. The bias of estimator T_4 is given by

$$B(T_4) = M_3\bar{X}\left[\frac{q\,(q+1)}{2}C_Y^2 - q\rho_{XY}C_XC_Y\right] \quad (4.10)$$

iii. The mean square error of T_4 is given by

$$M(T_4) = \bar{X}^2[M_1C_X^2 + q^2M_3C_Y^2 - 2qM_3\rho_{XY}C_XC_Y] \quad (4.11)$$

iv. The minimum mean squared error of T_4 is given by

$$M(T_4)_{min} = M_1S_X^2 - M_3\rho_{XY}^2\,S_X^2 \quad (4.12)$$

for the optimum value of $q = \rho_{XY}\frac{C_X}{C_Y}$

PROOFS:

i. $T_4 = \bar{x}_{..}^r\left(\frac{\bar{y}_{..}^n}{\bar{y}_{..}^r}\right)^q$

$= \bar{X}\,(1 + \varepsilon)(1+\eta)^q(1+\delta)^{-q}$ [by using first order approximation]

$= \bar{X}\left[1 + \varepsilon + q\eta - q\delta + q\varepsilon\eta - q\varepsilon\delta - q^2\delta\eta + \frac{q\,(q+1)}{2}\,\eta^2 + \frac{q\,(q+1)}{2}\,\delta^2\right]$

ii. $B(T_4) = E(T_4 - \bar{X})$

$$= \bar{X}\, E\Big(\varepsilon + q\eta - q\delta + q\varepsilon\eta - q\varepsilon\delta - q^2\delta\eta + \tfrac{q\,(q+1)}{2}\,\eta^2 + \tfrac{q\,(q+1)}{2}\,\delta^2\Big)$$

$$= M_3\bar{X}\left[\tfrac{q\,(q+1)}{2}C_Y^2 - q\rho_{XY}C_XC_Y\right]$$

iii. $M(T_4) = E(T_3 - \bar{X})^2 = \bar{X}^2\, E\Big[q^2\delta\eta + \tfrac{q\,(q+1)}{2}\,\eta^2 + \tfrac{q\,(q+1)}{2}\,\delta^2\Big]$

$$= \bar{X}^2[M_1C_X^2 + q^2M_3C_Y^2 - 2qM_3\rho_{XY}C_XC_Y]$$

iv. Minimum m.s.e. occurs when

$$\frac{d}{dq}M(T_4) = 0$$

we get optimum value of $q = \rho_{XY}\frac{C_X}{C_Y}$, for which m.s.e. becomes

$$M(T_4)_{min} = M_1S_X^2 - M_3\rho_{XY}^2\,S_X^2$$

Theorem 4.5:

i. The estimator T_5 in the relation of ε, η and δ is given by up to first order approximation is

$$T_5 = \bar{X}\left[1 + \varepsilon - s\delta - s\varepsilon\delta + \tfrac{s\,(s+1)}{2}\,\delta^2\right] \qquad (4.13)$$

ii. The bias of estimator T_5 is given by

$$B(T_5) = M_1\bar{X}\left[\tfrac{s\,(s+1)}{2}C_Y^2 - s\rho_{XY}C_XC_Y\right] \qquad (4.14)$$

iii. The mean square error of T_5 is given by

$$M(T_5) = M_1\bar{X}^2(C_X^2 + s^2\,C_Y^2 - 2s\rho_{XY}C_XC_Y) \qquad (4.15)$$

iv. The minimum mean squared error of T_5 is given by

$$M(T_5)_{min} = M_1S_X^2 - (1 - \rho_{XY}^2) \qquad (4.16)$$

for the optimum value of s is given by

$$s = \rho_{XY}\frac{C_X}{C_Y}$$

PROOFS:

i. $T_5 = \bar{x}_{..}^{r}\left(\frac{\bar{Y}}{\bar{y}_{..}^{r}}\right)^s$

$= \bar{X}\,(1 + \varepsilon)(1+\delta)^{-s}$[by using first order approximation]

$= \bar{X}\left[1 + \varepsilon - s\delta - s\varepsilon\delta + \tfrac{s\,(s+1)}{2}\,\delta^2\right]$ [ignoring higher terms of order n^{-1}]

ii. $B(T_5) = E(T_5 - \bar{X}) = \bar{X}\, E\Big(\varepsilon - s\eta - s\varepsilon\eta + \tfrac{s\,(s+1)}{2}\,\eta^2\Big)$

$$= M_1\bar{X}\left[\tfrac{s\,(s+1)}{2}C_Y^2 - s\rho_{XY}C_XC_Y\right]$$

iii. $M(T_5) = E(T_5 - \bar{X})^2 = \bar{X}^2\, E[\varepsilon - s\eta^2]$

$$= \bar{X}^2\,[E(\varepsilon)^2 - 2s\,E(\varepsilon\eta) + sE(\eta^2)]$$

$$= M_1\bar{X}^2(C_Y^2 + s^2\,C_Y^2 - 2s\rho_{XY}C_XC_Y)$$

iv. Minimum m.s.e. occurs when

$$\frac{d}{ds} M(T_5) = 0$$

for the optimum value of $s = \rho_{XY} \frac{C_X}{C_Y}$, we get m.s.e. as

$$M(T_5)_{min} = M_1 S_X^2 - (1 - \rho_{XY}^2)$$

5. Comparison

[1]: The minimum mean squared error of proposed estimator T_3 is

$$M(T_3)_{\text{min.}} = M_1 S_X^2 - M_2 \rho_{XY}^2 S_X^2$$

and the minimum mean squared error of estimator T_4 is

$$M(T_4)_{min.} = M_1 S_X^2 - M_3 \rho_{XY}^2 S_X^2$$

Now, let $D_1 = [M(T_3)_{\text{min.}} - M(T_4)_{\text{min.}}]$ (5.1)

$$= \rho_{XY}^2 S_X^2 (M_3 - M_2)$$

So, T_4 is better than T_3 if $D_1 > 0$

i.e. $\rho_{XY}^2 S_X^2 (M_3 - M_2) > 0 \Rightarrow \rho_{XY}^2 S_X^2 \left(\frac{1}{r} - \frac{2}{n}\right) > 0$

$$\Rightarrow r < \frac{n}{2}$$

Hence if the above condition holds , therefore T_4 is better than T_3 .

[2]: Let, $D_2 = [M(T_3)_{\text{min.}} - M(T_5)_{\text{min.}}] > 0$ (5.2)

$$= M_1 S_X^2 - M_2 \rho_{XY}^2 S_X^2 - M_1 S_X^2 - (1 - \rho_{XY}^2) > 0$$

$$= \rho_{XY}^2 S_X^2 (M_1 - M_2) > 0$$

Since $D_2 > 0$, therefore T_5 is always better than T_3.

[3]: $D_3 = [M(T_4)_{\text{min.}} - M(T_5)_{\text{min.}}] > 0$

$$= M_1 S_X^2 - M_3 \rho_{XY}^2 S_X^2 - M_1 S_X^2 - (1 - \rho_{XY}^2) > 0$$

$$= M_2 \rho_{XY}^2 S_X^2 \quad (5.3)$$

Since $D_3 > 0$, therefore T_5 is always better than T_4.

6. Empirical Study

The relative comparison among the estimators is given using a population generated the values of main variable (X) and auxiliary variable (Y) with size $N = 400$. The data for this illustration has been taken from the Appendix A. The summary of the population is:

$\bar{X} = 1.2275$; $\bar{Y} = .56500$; $S_x^2 = 12.6791$; $S_y^2 = 3.79790$; $\rho_{xy} = 0.80710$; $C_y = 3.44920$; $C_x = 2.9008$ and $p = q = s = \rho_{xy} \frac{C_x}{C_y} = 0.67877$

Table 6.1

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	2	24	5	0	0	0	0	0	0	0	0	0	0	0
0	0	1	22	5	4	5	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	2	0	4	8	0	0	0	0	0	33	0	0	0	27	0	0
0	0	0	0	0	0	0	0	0	0	0	0	7	6	7	1	0	5	0	0
0	0	0	0	0	0	0	21	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	4	0	5	7	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	5	7	0	7	7	6	3	0	0	0	0	0	0
0	0	0	5	4	3	0	5	8	4	5	1	0	5	0	0	0	0	0	0
0	7	65	0	4	5	0	9	0	0	0	0	0	3	1	0	0	0	0	0
0	1	4	5	0	7	3	3	0	0	0	0	0	0	0	0	0	0	0	0
0	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	21
0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Population of study variable (X)

Table 6.2

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	11	3	0	0	0	0	0	0	0	0	0	0	0
0	0	0	11	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	2	4	0	0	0	0	0	12	0	0	0	15	0	0
0	0	0	0	0	0	0	0	0	0	0	0	3	2	3	0	0	2	0	0
0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	2	0	2	3	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	2	2	0	3	3	2	1	0	0	0	0	0	0
0	0	0	2	2	1	0	2	3	2	2	0	0	2	0	0	0	0	0	0
0	3	18	0	2	2	0	4	0	0	0	0	0	1	0	0	0	0	0	0
0	0	2	2	0	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	12
0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Population of auxiliary variable variable (Y)

The bias and m.s.e. of the suggested estimator are calculated on 25,000 repeated samples selected by SRSWOR from population N = 400. Using a random sample of size n = 20 and f = 0.05 table 6.1 provides the results of the calculations. The simulation technique consists of the following steps:

1. Draw a random sample of size n = 20 from the population of N = 400 by SRSWOR.
2. Dropout 8 units randomly from each sample corresponding to X.
3. Compute and impute the dropped units of x with the help of proposed and available methods.
4. Repeat the above steps 30,000 times, which provides multiple sample based estimates.

5. Bias of $\widehat{y_i}$ is obtained by

$$B(\hat{x}) = \frac{1}{30,000} \sum_{i=1}^{30,000} [(\hat{x_i}) - \bar{X}]$$

6. M.S.E. of $\widehat{y_i}$ is obtained by

$$M(\hat{x}) = \frac{1}{30,000} \sum_{i=1}^{30,000} [(\hat{x}) - \bar{X}]^2$$

Table 6.3: Bias and M.S.E. of the estimator under study based on simulation

Estimator	Bias	MSE	Efficiency
T_1	4.315049	20.98326	1
T_2	0.117702	2.46123	8.5255
T_3	-0.61652	0.395419	53.0658
T_4	7.607369	63.4381	0.330767
T_5	-0.11	0.04082	514.0423

7. Conclusion:

From the study based on the results in table 6.1 we find that the proposed estimators have higher efficiency than those of mean method and ratio method. Moreover, we notice that the proposed estimators namely T_3, T_4 and T_5 we observe that the estimator T_5 is more efficient than the estimators T_3 and T_4 with less bias and less M.S.E. Hence we can say that the estimator T_5 is better than all the other estimators.

References

[1] Borkowski, J.J and Turk, P. (2013): *Adaptive cluster sampling,* International conference on Applied Statistics.

[2] Chutiman, N**.** (2013)**:** *Adaptive cluster sampling using auxiliary variable*, Journal of Mathematics and Statistics, 9(3), 249-255.

[3] Chutiman, N. Chiangpradit, M. (2014): *Ratio estimator in adaptive cluster sampling without replacement of networks.* J. Prob. Stat., 1-6, Article ID 726398 .

[4] Gattone, S.A. Mohaved, E., Dryver, A.L. and Munich, R.T. (2016): *Adaptive cluster sampling for negatively correlated data,* Environmetrics, 27, E103-E113.

[5] Latpate, R.V. and Kshirsagar, J.K. (2018a): *Negative adaptive cluster sampling, Model Assisted Statistics and Applications*, IOS Press.

[6] Latpate, R.V. and Kshirsagar, J.K. (2018b): *Two stage negative adaptive cluster sampling*, Communications in Mathematics and Statistics, (Springer) DOI: 10.1007/540304-018-0151-Z.

[7] Latpate, R.V. and Kshirsagar, J.K. (2018c): *Adaptive sampling for the improving sample performance*, Ph.D. Thesis, Submitted to Savitri Bai Phule_Pune University, Pune.

[8] Latpate, R.V. and Kshirsagar, J.K. (2018d): *Two stage inverse adaptive cluster sampling with stopping rule depends upon the size of cluster, Sankhya, Series B*, DOI.ORG/10.1007/513571-018-0177-y.

[9] Latpate, R.V. and Kshirsagar, J.K. (2018e): *Sample size considerations in the adaptive cluster sampling*, Bulletin of Marathwada Mathematical Society, Vol. 19(1), 32-41.

[10] Medina, M.H.F., Thompson, S.K. (2004): *Adaptive cluster double sampling.* Biometrika, vol. 91, No. 4, pp. 877-891.

[11] Narkhede, V., Latpate, R.V. and Kshirsagar, J.K. (2019)**:** *Review of adaptive sampling designs*, International Journal of Research and Analytical Reviews (IJRAR), 6(1), 35-41.

[12] Shukla, D. and Thakur N.S. (2008): *Estimation of mean with imputation of missing data using factor-type estimator,* Statistics in Transition, 9(1), p.33-48.

[13] Smith, D.R., Villella, R.F. and Lemarie (2003): *Application of adaptive cluster sampling to low density populations of freshwater mussels,* Environmental and Ecological Statistics, 10, 7-15.

[14] Singh, H.P. and Yadav**, A.** (2019)**:** *A class of estimators for estimating the population mean and variance using auxiliary information under adaptive cluster sampling in sample surveys*, Bulletin of Pure and Applied Sciences, 38E(1) (Math and Stat.), 176-192.

[15] Thakur, N. S. and Shukla D. (2022): *Missing data estimation based on the chaining technique in survey sampling,* Statistics in transition, Vol.23, No. 4, pp. 91-111, DOI 10.2478/stattrans-2022-0044.

[16] Thakur, N. S., Yadav, K. and Pathak, S. (2011): *Estimation of mean in presence of missing data under two-phase sampling scheme,* Journal of Reliability and Statistical studies, 4(2), 93-104p.

[17] Thakur, N.S., Yadav, K. and Pathak, S. (2014): *Estimation of mean with imputation of missing data in stratified random sampling,* Research & Reviews*:* A Journal of Statistics, Vol. 3, Issue 1, 17-30, ISSN: 2278 – 2273.

[18] Thakur, N.S. Yadav, K. and Singh, R. (2016): *A family of factor-type estimators for estimation of population* variance, Journal of Reliability and Statistical Studies, Vol. 9, Issue 2, 21-31, ISSN: 0974-8024 (print), 2229-5666 (online).

[19] Turk, P. and Borkowski J.J. (2005): *A review of adaptive cluster sampling,* Environmental and Ecological Statistics, 12, 55-94.

[20] Yadav, S.K., Mishra, S., S.S. and Chutiman, N. (2016): *Improved ratio estimators of population mean in adaptive cluster sampling,* Journal of Statistics Applications and Probability Letters, 3(1), 1-6.