Journal of Rajasthan Academy of Physical Sciences
ISSN: 0972-6306; URL: http://raops.org.in
Vol.24, No.1&2, January-June, 2025, 93-107

WEIGHTED PPS RATIO ESTIMATOR FOR POPULATION ESTIMATION WITH INCOMPLETE SAMPLING FRAMES

Jyoti[1*], Sarla Pareek[2] and **P.C. Gupta[3]**
[1,2]Department of Mathematics and Statistics, Banasthali Vidyapith, Rajasthan.
[3]Professor (Rtd.), Veer Narmad South Gujarat University, Gujarat.
Email: jyotidev626@gmail.com

Abstract: When conducting a sample survey, it is important to have a complete list of all units to be sampled, known as a sampling frame. However, sometimes this list may be incomplete, making it challenging to estimate the population's characteristics. This paper presents a new weighted PPS Ratio estimator for the population mean, which works for both complete and incomplete frame using Probability Proportional to Size with replacement (PPSWR) method. The paper calculates the bias and mean square error (MSE) of the estimator and discusses how to determine the best sample size and retainment factor with a cost-effective method. Using simulated data, the paper shows that this new estimator is more accurate and efficient as compared to the existing ones.

Keywords: Bias, MSE, variance, efficiency, linear cost function, Lagrange's Multiplier technique.

1. Introduction

In every data-based study, firstly the target population for which inferences are to be drawn, is defined. Many times, due to practical difficulty, it is not possible, to investigate each and every unit of the population. In that situation, only sample survey is the ray of hope. To get valid conclusions, selection of appropriate sampling method and selection of appropriate estimators are very crucial decisions. In every sampling method, it's essential to have a complete list of all sampling units which known as the sampling frame. The completeness of sampling in selecting samples and conducting surveys is always desirable. The details of the sampling frame help decide the best sampling design. However, in practice, it's rare to have a perfectly complete sampling frame. The most common issue with sampling frames is that they are often incomplete. In such cases, the foremost questions about the decision of sample size and selection of sample arise. Hansen and Hurwitz [5] were the first one to address this issue.

In many countries like India, the issue of incomplete frames in the context of data collection and sampling is a common challenge. This problem often arises due to factors such as limited technological resources and inadequate systems for data management and recording. In the first edition of his book, Yates [15] highlighted

incompleteness as a fundamental weakness of sampling frames, emphasizing the importance of addressing this issue to ensure the accuracy and reliability of statistical sampling and data collection in such contexts. Hartley [8] recommended using two or more frames to address the issue of incomplete frames, ensuring that the entire population is included. He also explained how to calculate the best sample sizes from these frames while considering the associated costs.

Seal [11] examined the use of outdated sampling frames in large-scale surveys. He viewed population changes as a continuous stochastic process and proposed employing successive frames to account for these changes when making estimates from incomplete data. Hansen to find data on missing units, Hansen and Hurwitz [5, 6] introduced a method called the predecessor-successor method. It is assumed that we can arrange units geographically so that for any given unit, we can easily find its successor by being included in the sample. However, this makes the sample size a random variable and increases the amount of work involved.

A significant contribution to the theory of estimation with incomplete frame was made by Agarwal and Gupta [1] by conducting a comprehensive review of the problem. They discussed some suitable estimation procedures with incomplete sampling frame under simple random sampling with and without replacement. Agarwal and Gupta [2] expanded the problem into a two-stage sampling design, where the incomplete frame corresponds to the second-stage unit. To estimate the total number of missing units in the sampling frame and the overall total for the variable being studied, Singh [13] introduced a method called the predecessor-successor method. He provided a mathematical way to use this method for estimating:

- The total number of units from the target population that are missing from the frame.
- The overall total of the characteristic being examined for the target population.

Singh [12] suggested a method for estimating when sampling from an incomplete frame, which may miss some units of the target population and include extra units that don't belong. He recommended using the predecessor-successor method, assuming that the units can be ordered geographically. The rules for ordering are set up so that for any unit, its successor can be found by following a specific travel path. Gupta [3] gave a detailed review of the problem of non-sampling error in general and that of randomized response error in particular. Gupta et al. [4] and Joshi et al. [9] proposed a weighted product and ratio estimator of population mean in case of incomplete sampling frames. In a recent development, Jyoti et al. [10] proposed a Hansen Hurwitz estimator that utilizes PPS sampling to address incomplete sampling frame situations.

This paper aims to contribute significantly to the literature by introducing an PPS Ratio estimator for the population mean that accounts for the incompleteness of the sampling frame. To address this issue, we consider two different cases. In Case I, we estimate the non-included units and also obtained the estimate of the target population. In Case II, we proposed a weighted PPS Ratio estimator which is a combination based on sampled units drawn from the units of complete frame and sample mean of the units sub-sampled from

the units occurring between the selected units and next to it. The problem is formulated as a single-objective optimization with a linear cost function and solved using the Lagrange multiplier technique. Additionally, a numerical illustration is provided using a simulated dataset to demonstrate the approach.

2. Estimation of Parameters by Predecessor-Successor Method

Let N represent the total number of units in the target population. Among these, let N_1 denote the number of units included in the sampling frame. Additionally, let N_2 (where N_2= N–N_1) represent the number of units that are also part of the target population but not listed in the sampling frame. But, its size (N_2) and list of such sampling units are unknown. How to estimate N_2 and doing survey from these identifying units is the challenge?

Further, let unidentified N_2 units lie among identified N_1 units. Let M represent the number of units from the target population that are missing from the frame, where the target population comprises N_1+ N_2 = N units. It is well understood that the N_2 units that no longer exist cannot be identified and, therefore, cannot be removed from the frame. In such situations, the Predecessor-Successor method proposed by Hansen et al. [7] can be employed to gather information about the units excluded from the frame.

The method relies on the assumption that a geographic ordering of units can be established, and the ordering rules ensure that, for any given unit in the population, its successor can be identified by following a defined travel path. This approach ensures that a unit missing from the frame (the successor) has the same probability of selection in the sample as its preceding unit listed in the frame. Consequently, the M units excluded from the frame may be located within N_1 gaps, arranged as follows:

U_1.....M_1; U_2…….M_2; U_3…...M_3; ……;U_{N_1}…….M_{N_1}

where, M_i is the number of non-included units in the i^{th} gap i.e. between i^{th} and $(i+1)^{th}$ unit.

Thus, $\sum_{i=1}^{N_1} M_i = M$; $0 \leq M_i \leq M$

Let Y: Population totals of the characteristic values for units in the entire target population.

Y_1: Population totals of the characteristic values for units in the frame.

Y_2: Population totals of the characteristic values for units not in the frame.

y_{n_1}: Totals for the characteristics under study obtained from samples of n_1 drawn from N_1.

y_{n_2}: Totals for the characteristics under study obtained from samples of n_2 drawn from N_2.

$y_{n_2'}$: Totals for the characteristics under study obtained from samples of n_2' drawn from n_2.

X: Auxiliary information for the population totals of the characteristic values for units in the entire population.

X_1: Auxiliary information for the population totals of the characteristic values for units in the frame.

X_2: Auxiliary information for the population totals of the characteristic values for units not in the frame.

$\bar{y}_{N_1} = \sum_{i=1}^{N_1} \frac{y_i}{N_1 P_i}$: mean for y of the N_1 units of the target population included in the frame.

$\bar{x}_{N_1} = \sum_{i=1}^{N_1} \frac{x_i}{N_1 P_i}$: mean for x of the N_1 units of the target population included in the frame.

$\bar{y}_{N_2} = \sum_{i=1}^{N_2} \frac{y_i}{N_2}$: mean for y of the N_2 units of the target population not included in the frame.

$P_i = \frac{x_i}{\sum x_i}$ i.e. P_i is proportional to x_i.

$r_i = \frac{y_i}{x_i}$: ratio of population total of y to the population total of x.

$R_{N_1} = \frac{\bar{Y}_{N_1}}{\bar{X}_{N_1}}$: ratio of the population mean of Y_{N_1} to the population mean X_{N_1}.

$r_{n_1} = \frac{\bar{y}_{n_1}}{\bar{x}_{n_1}}$: ratio of the sample mean of y_{n_1} to the sample mean x_{n_1}.

$R_{n_2} = \frac{\bar{y}_{n_2}}{\bar{x}_{n_2}}$: ratio of the sample mean of y_{n_2} to the sample mean x_{n_2}.

When conducting large surveys, two situations can occur with units not included in the survey frame:

- If the non-included units act similarly to the included ones.
- If the non-included units act differently from the included ones.

We will explain the estimation methods for each of these situations in the next sections.

Case I. Estimation procedure when the non-included units behave as included units:

Select a sample of size n_1 from N_1 by PPSWR with selection probabilities $P_i = \frac{x_i}{\sum x_i}$ (i = 1, 2, …, N_1) where X_i is the size of the i^{th} units in terms of some viable measurement. Let's assume that for the chosen units, the value being studied is y_{N_1}, and the sample values are y_{n_1}. All units in the gap between the selected ones, which are not part of the main frame, are automatically included in the sample with the same probability as the included units, $P_i = \frac{x_i}{\sum x_i}$. The units not included between the selected ones are M_i (i= 1, 2, …, n_1). Therefore, the estimated average number of non-included units between any two included ones is:

$$\bar{m} = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\mathrm{M_i}}{N_1 P_i} \quad (1)$$

$\bar{m}$ provided an unbiased estimator of $\bar{M} = \frac{1}{N_1}\sum_{i=1}^{N_1} M_i$

Whereas, $\bar{\mathrm{M}}$ represents the average number of units not included between any two included units in the population.

So, the total number of non-included units in the target population can be estimated as $N_1\bar{m}$. The variance of $\bar{\mathrm{M}}$ is calculated as:

$$\mathrm{v}\,(\bar{m}) = \frac{1}{n_1}\sum_{i=1}^{N_1} P_i\left[\frac{\mathrm{M}_i}{N_1 P_i} - \bar{M}\right]^2 \quad (2)$$

The estimate of variance of $\bar{m}$ is:

$$v\,(\hat{\bar{m}}) = \frac{1}{n_1(n_1-1)}\sum_{i=1}^{n_1}\left(\frac{{\mathrm{M}_i}^2}{N_1 P_i} - \bar{M}^2\right)$$

Thus, an estimate of target population is given as:

$$\hat{T}_1 = N_1(1+\bar{m})\bar{y}_{n_1} \quad (3)$$

where, $\bar{y}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} r_i\,\bar{x}_{N_1}$ is the sample mean of the n_1 included units.

and $\bar{m} = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\mathrm{M}_i}{N_1 P_i}$

The variance of T_1 is given by:

$$V\left(\hat{T}_1\right) = V\left[N_1(1+\bar{m})\bar{y}_{n_1}\right]$$

$$= N_1^2\left[V\left(\bar{y}_{n_1}\right) + V\left(\bar{m}\bar{y}_{n_1}\right) + 2\,COV\left(\bar{y}_{n_1}, \bar{m}\bar{y}_{n_1}\right)\right] \quad (4)$$

where, $V\left(\bar{y}_{n_1}\right) = \left(\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{y_i}{x_i}\bar{x}_{N_1}\right)$

$$= \frac{{\bar{x}_{N_1}}^2}{n_1}\sum_{i=1}^{N_1} P_i\left(r_i - \bar{r}_{N_1}\right)^2$$

$$V\left(\bar{m}\bar{y}_{n_1}\right) = V(\bar{m})\mathrm{V}\,(\bar{y}_{n_1}) + \mathrm{V}\,(\bar{m})\left[E\left(\bar{y}_{n_1}\right)\right]^2 + \mathrm{V}\,(\bar{y}_{n_1})[E(\bar{m})]^2$$

$$= V(\bar{m})\mathrm{V}\,(\bar{y}_{n_1}) + {\bar{y}_{N_1}}^2 V(\bar{m}) + \bar{M}^2\mathrm{V}\,(\bar{y}_{n_1})$$

and, $COV\left(\bar{y}_{n_1}, \bar{m}\bar{y}_{n_1}\right) = E(\bar{m})\mathrm{V}\,(\bar{y}_{n_1})$

$$= \bar{M}\mathrm{V}\,(\bar{y}_{n_1})$$

∴ From eq. (4), we get:

$$V\left(\hat{T}_1\right) = N_1^2\left[(1+\bar{M}^2)V\left(\bar{y}_{n_1}\right) + V(\bar{m})\left\{{\bar{y}_{N_1}}^2 + V\left(\bar{y}_{n_1}\right)\right\}\right] \quad (5)$$

Now, Estimate of $V(T_1)$ is:

$$\mathrm{V}(\hat{T}_1) = N_1^2\left[(1+\bar{M}^2)V(\hat{\bar{y}}_{n_1}) + V(\hat{\bar{m}})\{\bar{y}_{N_1}{}^2 + V(\hat{\bar{y}}_{n_1})\}\right] \quad (6)$$

where, $V(\hat{\bar{y}}_{n_1}) = \dfrac{\bar{x}_{N_1}{}^2}{n_1(n_1-1)}\sum_{i=1}^{n_1} P_i\left(r_i - \bar{r}_{n_1}\right)^2$

and, $\mathrm{V}(\hat{\bar{\mathrm{m}}}) = \frac{1}{\mathrm{n_1(n_1-1)}}\sum_{\mathrm{i=1}}^{\mathrm{n_1}}\left(\frac{\mathrm{M_i}^2}{\mathrm{N_1P_i}} - \bar{\mathrm{M}}^2\right)$

The above procedure gives a simple estimate for the total number of non-included units ($\mathrm{N}\bar{m}$), to estimate the average of the required characteristic for both included and non-included units ($\hat{\bar{y}}_{n_1}$), and to estimate the total of the target population ($\hat{T}_1$), assuming the non-included units behave similarly to the included ones.

Case II. Estimation procedure when the non-included units do not behave as included units:

Choose a sample of size n using the PPSWR method from a population with N =$N_1 + N_2$, where N_1 units are included in the sampling frame, and N_2 units are not included. For the selected units, record the value of the characteristic being studied, say y_n. Since the value of N_2 is completely unknown, we applied the subsampling technique outlined by Hansen and Hurwitz [6].

Now, let a sample of n units is selected which consist of two parts:

- First, a sample of n_1 units is selected from the available list of units using the Probability Proportional to Size with Replacement (PPSWR) method.
- Next, create a new list of units that fall between the selected units and those immediately after them. Let the total number of these non-included units be n_2, identified using the Predecessor-Successor method. From these n_2 units, choose n_2' units with PPSWR method using the Hansen-Hurwitz technique, considering n = $n_1 + n_2$

Suppose, $\bar{y}_{n_1}$ and $\bar{y}'_{n_2}$ are the sample means for the character under study.

Also, $w_1 = \frac{n_1}{n}$, $w_2 = \frac{n_2}{n}$, $k = \frac{n_2}{n_2'}$

where, w_1, w_2 are the weights of the sample and k is the retainment factor.

Also, $W_1 = \frac{N_1}{N}$, $W_2 = \frac{N_2}{N}$ are the respective population weights.

Since, our target population consist of two frames, one consists of those units which are included in the frame and other which are non-included.

Now, the estimate of N_2 can be given as:

$$\hat{N}_2 = \frac{n_2}{n}N \quad (7)$$

(i) Proposed Estimator:

The proposed estimator for estimating the population mean in the presence of incomplete sampling case is given by Unal and Kadilar [14]:

$$t_R = w_1 \frac{\bar{y}_{n_1}}{\bar{x}_{n_1}} \bar{X}_{N_1} + w_2 \frac{\bar{y}_{n_2'}}{\bar{x}_{n_2'}} \bar{x}_{n_2} \tag{8}$$

where, $w_1 = \frac{n_1}{n}$, $w_2 = \frac{n_2}{n}$ are the weights associated with complete and incomplete frames.

To obtain the Bias (t_R) and MSE (t_R), the notations are used under the Case II as follows:

$$\bar{y}_{n_1} = \bar{\mathrm{Y}}_{\mathrm{N}_1} + \bar{\varepsilon}_{n_1}\,,\ \bar{x}_{n_1} = \bar{\mathrm{X}}_{\mathrm{N}_1} + \bar{\varepsilon}'_{n_1} \tag{a}$$

$$\bar{y}_{n_2'} = \bar{\mathrm{y}}_{\mathrm{n}_2} + \bar{\varepsilon}_{n_2}\,,\ \bar{x}_{n_2'} = \bar{\mathrm{x}}_{\mathrm{n}_2} + \bar{\varepsilon}'_{n_2} \tag{b}$$

Such that:

$$\mathrm{E}\left(\bar{\varepsilon}_{n_1}\right) = \mathrm{E}\left(\bar{\varepsilon}'_{n_1}\right) = \mathrm{E}\left(\bar{\varepsilon}_{n_2}\right) = \mathrm{E}\left(\bar{\varepsilon}'_{n_2}\right) = 0 \tag{c}$$

$$\mathrm{E}\left(\bar{\varepsilon}^2_{n_1}\right) = \mathrm{V}\left(\bar{\mathrm{y}}_{n_1}\right);\ \mathrm{E}\left(\bar{\varepsilon}'^2_{n_1}\right) = \mathrm{V}\left(\bar{x}_{n_1}\right);\ \mathrm{E}\left(\bar{\varepsilon}^2_{n_2}\right) = \mathrm{V}\left(\bar{y}_{n_2}\right), \mathrm{E}\left(\bar{\varepsilon}'^2_{n_2}\right) = \mathrm{V}\left(\bar{x}_{n_2}\right) \tag{d}$$

$$\mathrm{E}\left(\bar{\varepsilon}_{n_1}\bar{\varepsilon}'_{n_1}\right) = Cov(\bar{y}_{n_1}, \bar{x}_{n_1});\ \mathrm{E}\left(\bar{\varepsilon}_{n_2}\bar{\varepsilon}'_{n_2}\right) = Cov(\bar{y}_{n_2}, \bar{x}_{n_2}) \tag{e}$$

(ii) Mean:

The estimation of the proposed estimator is as follows:

$$E_1E_2[t_R \mid n_1, n_2] = E_1E_2\left[w_1 \frac{\bar{y}_{n_1}}{\bar{x}_{n_1}} \bar{X}_{N_1} + w_2 \frac{\bar{y}_{n_2'}}{\bar{x}_{n_2'}} \bar{x}_{n_2}\right]$$

$$= E_1E_2\left[w_1 \frac{\bar{y}_{n_1}}{\bar{x}_{n_1}} \bar{X}_{N_1} \mid n_1, n_2\right] + E_1E_2\left[w_2 \frac{\bar{y}_{n_2'}}{\bar{x}_{n_2'}} \bar{x}_{n_2} \mid n_1, n_2\right]$$

Taking upto I^{st} order approximation, we obtain:

$$E(t_R) = -\left[\frac{W_1R_{N_1}}{n_1\bar{Y}_{N_1}} \sum_i^{N_1} P_i \left(\frac{y_i}{N_1P_i} - \bar{\mathrm{Y}}_{\mathrm{N}_1}\right)\left(\frac{x_i}{N_1P_i} - \bar{\mathrm{X}}_{\mathrm{N}_1}\right) + \frac{W_2R_{n_2}}{n_2\bar{y}_{n_2}} \sum_i^{n_2} P_i \left(\frac{y_i}{n_2P_i} - \bar{\mathrm{y}}_{\mathrm{n}_2}\right)\left(\frac{x_i}{n_2P_i} - \bar{\mathrm{x}}_{\mathrm{n}_2}\right)\right] + \bar{y}_N \# \tag{9}$$

Therefore, ratio estimator is a biased estimator. The Bias of the proposed estimator is:

$$\mathrm{Bias}(t_R) = -\left[\frac{W_1R_{N_1}}{n_1\bar{Y}_{N_1}} \sum_i^{N_1} P_i \left(\frac{y_i}{N_1P_i} - \bar{\mathrm{Y}}_{\mathrm{N}_1}\right)\left(\frac{x_i}{N_1P_i} - \bar{\mathrm{X}}_{\mathrm{N}_1}\right) + \frac{W_2R_{n_2}}{n_2\bar{y}_{n_2}} \sum_i^{n_2} P_i \left(\frac{y_i}{n_2P_i} - \bar{\mathrm{y}}_{\mathrm{n}_2}\right)\left(\frac{x_i}{n_2P_i} - \bar{\mathrm{x}}_{\mathrm{n}_2}\right)\right] \# \tag{10}$$

(iii) Variance:

Since, we know that:

$$\mathrm{V}(t_R) = V_1E_2(t_R) + E_1V_2(t_R) \tag{11}$$

Here,

$$V_1E_2(t_R) = V_1E_2\left(w_1\frac{\bar{y}_{n_1}}{\bar{x}_{n_1}}\bar{X}_{N_1} + w_2\frac{\bar{y}_{n_2'}}{\bar{x}_{n_2'}}\bar{x}_{n_2}\right)$$

Using the above relations ((a) to (e)) we have:

$$V_1E_2(t_R) = V_1\left(w_1\frac{\bar{\mathrm{Y}}_{\mathrm{N}_1}+\bar{\varepsilon}_{n_1}}{\bar{\mathrm{X}}_{\mathrm{N}_1}+\bar{\varepsilon}'_{n_1}}\bar{X}_{N_1} + w_2\frac{\bar{\mathrm{y}}_{\mathrm{n}_2}+\bar{\varepsilon}_{n_2}}{\bar{\mathrm{x}}_{\mathrm{n}_2}+\bar{\varepsilon}'_{n_2}}\bar{x}_{n_2}\right)$$

$$= V_1\left(w_1\frac{\bar{\mathrm{Y}}_{\mathrm{N}_1}+\bar{\varepsilon}_{n_1}}{\bar{\mathrm{X}}_{\mathrm{N}_1}+\bar{\varepsilon}'_{n_1}}\bar{X}_{N_1} + w_2\left\{\bar{Y}_{N_2} - \frac{R_{n_2}}{n_2\bar{y}_{n_2}}\sum_i^{n_2}P_i\left(\frac{y_i}{n_2P_i} - \bar{\mathrm{y}}_{\mathrm{n}_2}\right)\left(\frac{x_i}{n_2P_i} - \bar{\mathrm{x}}_{\mathrm{n}_2}\right)\right\}\right)$$

$$= \mathrm{V}\left(w_1\frac{\bar{\mathrm{Y}}_{\mathrm{N}_1}+\bar{\varepsilon}_{n_1}}{\bar{\mathrm{X}}_{\mathrm{N}_1}+\bar{\varepsilon}'_{n_1}}\bar{X}_{N_1}\right)$$

After simplifying the above term up to first order approximation, we have:

$$V_1E_2(t_R) = \mathrm{E}\left(\bar{\varepsilon}_{n_1}^2 + w_1^2\bar{\varepsilon}_{n_1}'^2R_{N_1}^2 - 2w_1\bar{\varepsilon}_{n_1}\bar{\varepsilon}'_{n_1}R_{N_1}\right)$$

$$= \mathrm{E}\left(\bar{\varepsilon}_{n_1}^2\right) + R_{N_1}^2\mathrm{E}\left(w_1^2\bar{\varepsilon}_{n_1}'^2\right) - 2R_{N_1}\mathrm{E}\left(w_1\bar{\varepsilon}_{n_1}\bar{\varepsilon}'_{n_1}\right)$$

$$= w_1^2V\left(\bar{y}_{n_1}\right) + w_1^2R_{N_1}^2\mathrm{V}\left(\bar{x}_{n_1}\right) - \frac{2W_1R_{N_1}Cov(\bar{y}_{n_1},\bar{x}_{n_1})}{\bar{\mathrm{Y}}_{N_1}} \tag{12}$$

and,

$$E_1V_2(t_R) = E_1\left[\frac{w_2^2}{n_2'}V\left(\frac{\bar{\mathrm{y}}_{\mathrm{n}_2}+\bar{\varepsilon}_{n_2}}{\bar{\mathrm{x}}_{\mathrm{n}_2}+\bar{\varepsilon}'_{n_2}}\bar{x}_{n_2}\right)\right]$$

$$= \frac{w_2^2k}{n_2(n_2-k)}\left[\begin{array}{c}\sum_i^{n_2'}P_i\left(\frac{y_i}{n_2P_i} - \bar{\mathrm{y}}_{n_2'}\right)^2 + \frac{R_{n_2}^2}{\bar{x}_{n_2}^2}\sum_i^{n_2'}P_i\left(\frac{x_i}{n_2P_i} - \bar{\mathrm{x}}_{n_2'}\right)^2 \\ - \frac{2R_{N_1}}{\bar{\mathrm{Y}}_{\mathrm{N}_1}}\sum_i^{N_1}P_i\left(\frac{y_i}{n_2P_i} - \bar{\mathrm{y}}_{n_2'}\right)\left(\frac{x_i}{n_2P_i} - \bar{\mathrm{x}}_{n_2'}\right)\end{array}\right] \tag{13}$$

Now, putting the value of eq. (12) & (13) in eq. (11), we get:

$$\mathrm{V}(t_R) = \frac{w_1^2}{n_1}\left[\sum_i^{N_1}P_i\left(\frac{y_i}{N_1P_i} - \bar{\mathrm{Y}}_{N_1}\right)^2 + \frac{R_{N_1}^2}{\bar{X}_{N_1}^2}\sum_i^{N_1}P_i\left(\frac{x_i}{N_1P_i} - \bar{\mathrm{X}}_{\mathrm{N}_1}\right)^2 - \frac{2R_{N_1}}{\bar{\mathrm{Y}}_{\mathrm{N}_1}}\sum_i^{N_1}P_i\left(\frac{y_i}{N_1P_i} - \bar{\mathrm{Y}}_{\mathrm{N}_1}\right)\left(\frac{x_i}{N_1P_i} - \bar{\mathrm{X}}_{\mathrm{N}_1}\right)\right] + \frac{w_2^2k}{n_2(n_2-k)}\left[\sum_i^{n_2'}P_i\left(\frac{y_i}{n_2P_i} - \bar{\mathrm{y}}_{n_2'}\right)^2 + \frac{R_{n_2}^2}{\bar{x}_{n_2}^2}\sum_i^{n_2'}P_i\left(\frac{x_i}{n_2P_i} - \bar{\mathrm{x}}_{n_2'}\right)^2 - \frac{2R_{N_1}}{\bar{\mathrm{Y}}_{\mathrm{N}_1}}\sum_i^{N_1}P_i\left(\frac{y_i}{n_2P_i} - \bar{\mathrm{y}}_{n_2'}\right)\left(\frac{x_i}{n_2P_i} - \bar{\mathrm{x}}_{n_2'}\right)\right]$$

$$\therefore \mathrm{V}(t_R) = w_1^2A + w_2^2B \tag{14}$$

where,

$$\mathrm{A}=\frac{1}{n_1}\left[\sum_i^{N_1}P_i\left(\frac{y_i}{N_1P_i} - \bar{\mathrm{Y}}_{N_1}\right)^2 + \frac{R_{N_1}^2}{\bar{X}_{N_1}^2}\sum_i^{N_1}P_i\left(\frac{x_i}{N_1P_i} - \bar{\mathrm{X}}_{\mathrm{N}_1}\right)^2 - \frac{2R_{N_1}}{\bar{\mathrm{Y}}_{\mathrm{N}_1}}\sum_i^{N_1}P_i\left(\frac{y_i}{N_1P_i} - \bar{\mathrm{Y}}_{\mathrm{N}_1}\right)\left(\frac{x_i}{N_1P_i} - \bar{\mathrm{X}}_{\mathrm{N}_1}\right)\right]$$

$$\mathrm{B}= \frac{k}{n_2(n_2-k)}\left[\sum_i^{n_2'} P_i\left(\frac{y_i}{n_2P_i}-\bar{\mathrm{y}}_{n_2'}\right)^2+\frac{R_{n_2}^2}{\bar{x}_{n_2}^2}\sum_i^{n_2'} P_i\left(\frac{x_i}{n_2P_i}-\bar{\mathrm{x}}_{n_2'}\right)^2-\frac{2R_{N_1}}{\bar{\mathrm{Y}}_{\mathrm{N}_1}}\sum_i^{N_1} P_i\left(\frac{y_i}{n_2P_i}-\bar{\mathrm{y}}_{n_2'}\right)\left(\frac{x_i}{n_2P_i}-\bar{\mathrm{x}}_{n_2'}\right)\right]$$

(iv) Mean Square Error (MSE):

$$\mathrm{MSE}(t_R)=w_1^2A+w_2^2\mathrm{B}+D^2 \tag{15}$$

where,

$$\mathrm{D}=-\left[\frac{W_1R_{N_1}}{n_1\bar{Y}_{N_1}}\sum_i^{N_1} P_i\left(\frac{y_i}{N_1P_i}-\bar{\mathrm{Y}}_{\mathrm{N}_1}\right)\left(\frac{x_i}{N_1P_i}-\bar{\mathrm{X}}_{\mathrm{N}_1}\right)+\frac{W_2R_{n_2}}{n_2\bar{y}_{n_2}}\sum_i^{n_2} P_i\left(\frac{y_i}{n_2P_i}-\bar{\mathrm{y}}_{\mathrm{n}_2}\right)\left(\frac{x_i}{n_2P_i}-\bar{\mathrm{x}}_{\mathrm{n}_2}\right)\right]$$

(v) Estimate of the Variance:

Hence, estimate of $\mathrm{V}(\bar{y}_{WR})$ is:

$$\mathrm{V}(\hat{t}_R)=\frac{w_1^2}{n_1}\left[\sum_i^N P_i\left(\frac{y_i}{N_1P_i}-\bar{\mathrm{Y}}_N\right)^2+\frac{R_{N_1}^2}{\bar{X}_{N_1}^2}\sum_i^N P_i\left(\frac{x_i}{N_1P_i}-\bar{\mathrm{X}}_N\right)^2-\frac{2R_{N_1}}{\bar{\mathrm{Y}}_{\mathrm{N}_1}}\sum_i^N P_i\left(\frac{y_i}{N_1P_i}-\bar{\mathrm{Y}}_N\right)\left(\frac{x_i}{N_1P_i}-\bar{\mathrm{X}}_N\right)\right]+\frac{w_2^2k}{n_2(n_2-k)}\left[\sum_i^{n_2} P_i\left(\frac{y_i}{n_2P_i}-\bar{\mathrm{y}}_{n_2}\right)^2+\frac{R_{n_2}^2}{\bar{x}_{n_2}^2}\sum_i^{n_2} P_i\left(\frac{x_i}{n_2P_i}-\bar{\mathrm{x}}_{n_2}\right)^2-\frac{2R_{N_1}}{\bar{\mathrm{Y}}_{\mathrm{N}_1}}\sum_i^{n_2} P_i\left(\frac{y_i}{n_2P_i}-\bar{\mathrm{y}}_{n_2}\right)\left(\frac{x_i}{n_2P_i}-\bar{\mathrm{x}}_{n_2}\right)\right] \tag{16}$$

3. Optimum Variance under some Condition

For obtaining the optimum variance, we need to obtain the optimum value of weights. So, we take a condition for weights as:

$$w_1+w_2=1$$
$$w_2=1-w_1$$

Putting w_2 as $(1-\mathrm{w}_1)$ in eq. (14) and finding its derivative w.r.t. w_2, we obtain:

$$\frac{\mathrm{dV}(t_R)}{\mathrm{dw}_2}=0$$

The optimum values of w_1 and w_2 obtained are:

$$(w_1)_{opt}=\frac{B}{\mathrm{A}+\mathrm{B}}$$

$$(w_2)_{opt}=\frac{\mathrm{A}}{\mathrm{A}+\mathrm{B}}$$

After substituting the values of w_1 and w_2 into equation (14), we obtain the optimum variance as:

$$V(t_R)_{opt}=\mathrm{B}-\frac{B^2}{A} \tag{17}$$

4. Cost Function

In every sample survey, costs are an unavoidable aspect. The total cost of conducting the survey depends on various factors such as overhead costs, travel costs, enumeration costs, and so on. However, these costs are bounded by certain limits. As a result, the objective is to find an estimator with minimum variance and fixed cost. Here, the linear cost function C is defined as follows:

$$C' = C_0 + C_1 n_1 + C_2 n_2$$
$$= C_0 + C_1 n_1 + C_2 n_2' k$$

Since, C' varies from sample to sample, we take the expected cost as:

$E(C')$= C

$$E(C)=C = C_0 + n\left(W_1 C_1 + \frac{W_2 C_2}{k}\right) \quad (18)$$

where, C_0: Cost of establishment
C_1: Evaluation cost of unit in the frame
C_2: Evaluation cost of unit not in the frame
$k = \frac{n_2}{n_2'}$: Constant retainment factor

For a fixed expected cost C_0, the problem of finding the optimum sample 'n' and retainment factor 'k' may be stated as the following:

Minimize $V = \mathrm{B} - \frac{B^2}{A}$ (19)
subject to $n\left(W_1 C_1 + \frac{W_2 C_2}{k}\right) \leq C_0$
and $n, \mathrm{k} \geq 0$

5. Lagrange's Multiplier (LM) Technique

The method of LM is applied to obtain optimal values of k and n. The Lagrange function is defined as follows:

$$L(n, \lambda, k) = V(t_R)_{\text{opt}} + \lambda(C - C_0)$$
$$= \mathrm{B} - \frac{B^2}{A} + \lambda\left[n\left(W_1 C_1 + \frac{W_2 C_2}{k}\right) - C_0\right] \quad (20)$$

where, λ is the Lagrange multiplier.

The necessary conditions for the solution of the problem: $\frac{dL}{dn} = 0$ which gives

$$-\frac{B^2}{A} + \lambda\left(W_1 C_1 + \frac{W_2 C_2}{k}\right) = 0 \quad (21)$$

Also, $\frac{dL}{d\lambda} = 0$

which gives

$$n\left(W_1C_1+\frac{W_2C_2}{k}\right)-C_0=0 \tag{22}$$

and, $\frac{dL}{dk}=0$

which gives

$$\frac{n_2'\mathrm{B}^2}{A}-\frac{\lambda nW_2C_2}{k^2}=0$$

$$\lambda=\frac{n_2\mathrm{kB}^2}{nW_2C_2A} \tag{23}$$

Substituting value of λ in eq. (21), we find:

$$k(\text{ opt. })=\frac{\mathrm{n}_1W_2C_2}{\mathrm{n}_2W_1C_1}$$

Substituting value of λ in eq. (22), we find:

$$n(\text{ opt. })=\sqrt{\frac{n_1C_0}{W_1C_1}}$$

6. Comparative Study with other existing Estimators

Several estimators have been proposed in the literature for estimating the population mean using the sub-sampling method. Table 1 lists these estimators along with their Mean Squared Error (MSE) equations, approximated to the first order, for Case II.

Table 1. List of estimators for Case II.

Vohita et al. [15]	$\bar{y}_{wr}=w_1\bar{y}_r+w_2\bar{y}_{n_2'}$	$\mathrm{MSE}(\bar{y}_{\mathrm{wr}})=\left(\frac{1-f}{n}\right)S_Y^2+W_1\left(\frac{1-f_1}{n}\right)R_1^2S_{X_1}^2-2W_1R_1\left(\frac{1-f_1}{n}\right)S_{Y_1X_1}+W_2\left(\frac{\mathrm{k}-1}{n}\right)S_{Y_2}^2+\left\{W_1\left(\frac{1}{n_1}-\frac{1}{N_1}\right)\frac{1}{\bar{X}_{N_1}}\left(R_1S_{X_1}^2-S_{X_1Y_1}\right)\right\}^2$
Jyoti et al. [6]	$\bar{y}_{PPSWR}=w_1\bar{y}_{HH}+w_2\bar{y}_{n_2'}$	$\mathrm{MSE}(\bar{y}_{PPSWR})=\frac{w_1}{\mathrm{n}}\sum_{i=1}^{N_1}P_1\left(\frac{Y_1}{N_1P_1}-Y\right)^2+w_2\frac{(k-1)}{n}S_{Y_2}^2$

Gupta et al. [1]	$\bar{y}_{AG} = w_1\bar{y}_{n_1} + w_2\bar{y}_{n_2'}$	$MSE(\bar{y}_{AG}) = \left(\frac{1-f}{n}\right)S_Y^2 + W_2\left(\frac{k-1}{n}\right)S_{Y_2}^2$
Vohita et al. [12]	$\bar{y}_{wp} = w_1\bar{y}_p + w_2\bar{y}_{n_2'}$	$\mathrm{MSE}(\bar{y}_{wr}) = W_1\left[\left(\frac{f_1}{f}\right)\left(\frac{1-f_1}{n}\right)\left[R_1^2S_{X_1}^2 + S_{Y_1}^2 - 2R_1S_{(XY)_1}\right]\right] + W_2\left[\left(\frac{f_2}{f}\right)\left(\frac{1-f_2}{n}\right)R_1^2 + \left(\frac{k-1}{n}\right)\right]S_{Y_2}^2 + \left\{W_1\left(\frac{1}{n_1} - \frac{1}{N_1}\right)\frac{S_{XY}}{\bar{X}_{N_1}} + \bar{Y}_N\right\}^2$

6.1 Efficiency Comparisons:

The proposed estimator t_R is compared with several other estimators listed in Table 1 to show how suitable it is for Case II. We compare the Mean Squared Error MSE (t_R) with the MSEs of the estimators in Table 1 and determine the efficiency conditions for Case II as follows:

- MSE (t_R) < MSE $(\bar{y}_{wr})$

$$[w_1^2A + w_2^2B + D^2] - \left\{\left(\frac{1-f}{n}\right)S_Y^2 + W_1\left(\frac{1-f_1}{n}\right)R_1^2S_{X_1}^2 - 2W_1R_1\left(\frac{1-f_1}{n}\right)S_{Y_1X_1} + W_2\left(\frac{k-1}{n}\right)S_{Y_2}^2\right\} + \left\{W_1\left(\frac{1}{n_1} - \frac{1}{N_1}\right)\frac{1}{\bar{X}_{N_1}}\left(R_1S_{X_1}^2 - S_{X_1Y_1}\right)\right\}^2 < 0 \quad (24)$$

- MSE (t_R) < MSE $(\bar{y}_{PPSWR})$

$$[w_1^2A + w_2^2B + D^2] - \left\{\frac{w_1}{n}\sum_{i=1}^{N_1} P_1\left(\frac{Y_1}{N_1P_1} - Y\right)^2 + w_2\frac{(k-1)}{n}S_{Y_2}^2\right\} < 0 \quad (25)$$

- MSE (t_R) < MSE $(\bar{y}_{AG})$

$$[w_1^2A + w_2^2B + D^2] - \left\{\left(\frac{1-f}{n}\right)S_Y^2 + W_2\left(\frac{k-1}{n}\right)S_{Y_2}^2\right\} < 0 \quad (26)$$

- MSE (t_R) < MSE $(\bar{y}_{wp})$

$$[w_1^2A + w_2^2B + D^2] - \left\{W_1\left[\left(\frac{f_1}{f}\right)\left(\frac{1-f_1}{n}\right)\left[R_1^2S_{X_1}^2 + S_{Y_1}^2 - R_1S_{(XY)_1}\right]\right] + W_2\left[\left(\frac{f_2}{f}\right)\left(\frac{1-f_2}{n}\right)R_1^2 + \left(\frac{k-1}{n}\right)\right]S_{Y_2}^2\right\} + \left\{W_1\left(\frac{1}{n_1} - \frac{1}{N_1}\right)\frac{S_{XY}}{\bar{X}_{N_1}} + \bar{Y}_N\right\}^2 < 0 \quad (27)$$

Under the conditions between eq. (24) - (27) for case II, we conclude from the condition results that the t_R estimator performs better than other estimators in the literature.

7. Simulation Study

We evaluate the performance of the proposed estimator t_R using a simulation study conducted in R software. In this Monte Carlo simulation, we consider a population of N=1000 observations, where 25% of the observations are not included in the sampling frame using different k values. The values for the auxiliary and study variables are generated from a multivariate normal distribution. For both Cases I and II, the dataset is assumed to follow a bivariate normal distribution with a mean of (7, 1), standard deviations of 0.1 and 10, and a correlation coefficient of 0.95, as used by Unal and Kadilar [14].

In Case I, Table 2 have been constructed which shows the estimate of the total number of non-included units ($\widehat{N}_2$), sample average number of non-included units each gap of included units of the frame ($\overline{m}$) and estimate of target population ($\hat{T}_1$).

Table 2. Obtained results of simulated data for study variable (Y) and auxiliary variable (X)

N_1	N_2	$\widehat{N}_2$	$\overline{m}$	V ($\overline{m}$)	S.E. ($\overline{m}$)	$\hat{T}_1$	V ($\hat{T}_1$)	S.E. ($\hat{T}_1$)
750	250	250	0.33324	0.000580	0.024087	8342.505	203565.3	451.182

In Case II, the MSE and PRE values for the proposed estimator and other existing estimators are shown in Table 3. The PRE values are calculated relative to the Agarwal and Gupta estimator ($\bar{y}_{AG}$), according to case II, respectively. Additionally, the Percent Relative Efficiencies (PREs) of the proposed estimator (t_R) and other existing estimators compared to the Agarwal and Gupta estimator ($\bar{y}_{AG}$) are determined using the PRE formula below:

$$\text{PRE}\ (\bar{y}_{**}) = \frac{MSE\ (\bar{y}_{AG})}{MSE(\bar{y}_{**})} \times 100$$

Table 3: MSE and PRE values of the t_R and compared estimators for case II under the simulation study.

	MSE of the Estimators				**PRE of the Estimators**			
	ρ_{XY}= 0.95				**ρ_{XY}= 0.95**			
Estimator	**k=2**	**k=3**	**k=4**	**k=5**	**k=2**	**k=3**	**k=4**	**k=5**
t_R	**0.050**	**0.069**	**0.088**	**0.108**	**1126**	**952.173**	**853.409**	**781.481**
$\bar{y}_{AG}$	0.563	0.657	0.751	0.844	100	100	100	100
$\bar{y}_{wp}$	3.412	3.507	3.602	3.697	16.50	18.73	20.84	22.82
$\bar{y}_{wr}$	0.405	0.500	0.595	0.690	139.01	131.40	126.21	122.31
$\bar{y}_{PPSWR}$	0.508	0.626	0.745	0.863	110.82	104.95	100.74	97.75

According to Table 3, the proposed estimator t_R is more efficient than compared estimators under Case II. The proposed estimator for each value has the minimum MSE value and maximum PRE value.

8. Conclusion

This paper aims to propose an estimator to accurately infer population parameters in case of incomplete sampling frames. For estimating the parameters, two cases i.e. Case I and II are discussed separately. In Case I, the estimate of the total number of non-included units ($\widehat{N}_2$) is calculated which is exactly equal to the N_2 i.e 250. Similarly, the estimate of the total population comes out to be 8342 which are not much different from the actual population of 7863.

In Case II, the Hansen–Hurwitz method is used to propose a ratio estimator with a linear cost function for estimating incomplete sampling frames. First, the proposed estimator is theoretically compared to existing estimators listed in Table 1 for relevant cases. Its performance is evaluated against other estimators under PPSWR using simulated data. The results suggest that the proposed estimator can be used under specific conditions outlined in equations (17) – (20) for Case II. Statistical properties like bias, mean squared error (MSE), and minimum MSE are analysed. Simulations show that the proposed estimator achieves the lowest MSE and highest PRE value compared to others in the incomplete sampling frame scenario, as shown in Table 3. Based on these findings, we recommend using the proposed estimator for incomplete sampling frame cases.

Acknowledgement: We would like to show our gratitude to Prof. P.C. Gupta for sharing his pearls of wisdom with us and helping us with the concepts.

References

[1] Agarwal, B. and Gupta, P. C. (2008). Estimation from incomplete sampling frames in case of simple random sampling. Model Assisted Statistics and Applications, **3**(2), 113-117.

[2] Agarwal, B. and Gupta, P. C. (2012). Estimation from incomplete sampling frame for twostage sampling design. Indian Journal of Statistical Application, **1**(2), 52-58.

[3] Gupta, P. C. (2004). Estimation from incomplete samples in finite population with special reference to randomized response technique. VNSG Univ. Journal, **2**, 133-139.

[4] Gupta, P. C., Joshi, L., Joshi, V. and Nagar, P. (2019). Weighted product estimation for incomplete sampling frames. Journal of Rajasthan Academy of Physical Sciences, **18**(3,4), 233-240.

[5] Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. The Annals of Mathematical Statistics, **14**(4), 333-362.

[6] Hansen, M. H. and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. Journal of the American Statistical Association, **41**(236), 517-529.

[7] Hansen, M. H., Hurwitz, W. N. and Jabine, T. B. (1963). The use of imperfect lists for probability- sampling at the united-states bureau of the census. Bulletin of the International Statistical Institute, **40**(1), 497-517.

[8] Hartley, H. O. (1962). Multiple frame surveys. In Proceedings of the Social Statistics Section, America Statistical Association, **19**(6), 203-206.

[9] Joshi, V., Nagar, P., Singh, A. K., and Gupta, P.C. (2021). Use of ratio method of estimation in incomplete frames. International Journal of Agricultural & Statistical Sciences, **17**(1), 545-549.

[10] Jyoti, Pareek, S. and Gupta, P. C. (2023). Estimation with incomplete frames using varying probability sampling with replacement. Model Assisted Statistics and Applications, **18**(1), 85-91.

[11] Seal, K. C. (1962). Use of out – dated frames in large scale sample surveys. Calcutta Statistical Association Bulletin, **11**(3), 68-84.

[12] Singh, R. (1989). Method of estimation for sampling from incomplete frames. Australian Journal of Statistics, **31**(2), 269-276.

[13] Singh, R. (1983). On the use of incomplete frames in sample surveys. Biometrical Journal, **25**(6), 545-549.

[14] Unal, C. and Kadilar, C. (2022). A new population mean estimator under non – response cases. Journal of Talibah University for Science, **16**(1), 111-119.

[15] Yates, F. (1948). Sampling Methods for Census and Surveys. London: Charles Griffin and Co.