Journal of Rajasthan Academy of Physical Sciences
ISSN: 0972-6306; URL: http://raops.org.in
Vol.21, No.3&4, July-December, 2022, 255-266

PARAMETER ESTIMATION IN CASE OF INCOMPLETE SAMPLING FRAMES: A SIMULATION APPROACH

Shilpa Yadav[1], Rashmi Bundel[1], Ajeet Kumar Singh[1] and Pankaj Nagar[2]
[1]Assistant Professor, Department of Statistics, University of Rajasthan, Jaipur
[2]Professor, Department of Statistics, University of Rajasthan, Jaipur.
Email: pnagar121@gmail.com (Corresponding Author)

Abstract: In this paper, the authors have dealt with incomplete sampling frames and tried to use a weighted estimator to get refined and improved estimates of the population characteristics and their standard errors for simple random sampling cases (with and without replacement schemes). Further, considering the cost restrictions for surveys, the optimum values of the sample sizes have been obtained using a suitable cost function that has not been taken up in the past for such problems. Finally, a Monte Carlo simulation technique is applied to numerically illustrate the efficacy and applicability of the technique in real-life situations.

Keywords: Imperfect frames, predecessor-successor method, non-sampling error, cost function, Monte Carlo simulation.

1. Introduction

The availability of a well-defined sampling frame is the most cardinal prerequisite for applying a probability sampling technique. The sampling frame is complete if units in the population are included only once in the frame and the repetition of items or irrelevant items are excluded from the study. The basic techniques of probability sampling, as discussed by Neyman [9], such as simple random sampling (SRS), stratified random sampling and systematic random sampling, presume the availability of such a sampling frame. However, it is not easy to have a perfect frame with a list that can be used for sampling in practice. Imperfection in sampling frames may arise due to four primary factors as discussed by Kish [8]. The omission of sampling units that should have been a part of the population, but are not included, leads to the first and most common problem, i.e. incompleteness in sampling frames, which adds to the non-sampling error. Even if the surveyor succeeds in preparing an accurate and complete frame that covers the entire population, by the time the survey starts, it becomes outdated and leads to an incomplete sampling frame problem. The second problem involves the clustering of sampling units which violates the condition of individual listing of the sampling units. Blank listing and duplication are other reasons for imperfection in the frames.

For example, in demographic surveys to study infant mortality rates, the list of infants born in the past year from hospital records is obtained. No matter how judiciously the frame is prepared, the frame would still be imperfect as many of the infant births and deaths would still not be recorded or registered in any hospital records. In medical surveys, there is rarely a situation when one can obtain a complete list of all the units desired for any study due to under-registration and incomplete patient information. Even in household surveys, real-time immigration and migration make it difficult to prepare a perfect sampling frame which is accurate and up to date. The problem of NRC in India and Banjara tribes in Rajasthan are evident examples of incomplete sampling frame problems.

The incomplete information obtained using an imperfect frame makes the full survey and its results ambiguous and misleading. Therefore, it is of vital importance to develop new techniques to deal with such situations. The problem of imperfect frames was first discussed by Hansen et al. [4], where they proposed the predecessor-successor method to gather information on the units which have been omitted in the sampling frame for some reason. The method assumes that the sampling units can be established in a geographical order such that given any one unit in the population, one can unambiguously determine its successor by following a well-defined path. This method provides an equal chance of selection to the units omitted from the frame; as a result, the workload increases because, in this case, the sample size becomes a random variable. For instance, there are four industrial units employing more than 100 employees registered with the registrar of industries in a district. But two of them start 3 new industries which have not been registered as their registration was not required as per the Industrial Act. In such cases, to study the socio-economic conditions, the original sampling frame would be incomplete, but the predecessor-successor method proposed can efficiently be used to gain a better estimate of population characteristics.

Hartley [6] proposed the use of two or more frames to overcome the problem of incompleteness in frames such that the combination of these frames provides us with a complete layout of the population when taken together. He has also obtained the optimum sample size to be taken from each frame under a suitable cost function. Seal [10] attempted to use outdated frames for estimating the total or average of characteristic under the study of a dynamic population at any given time. Later on, the problem was taken into consideration by Singh [12,13], where he used the predecessor-successor method for collecting information on the units which have been excluded in the sampling frame but are of vital importance for the study. Using the data collected, he proposed an estimator that would provide the improved estimates of the population characteristic and its standard error. Agarwal and Gupta [1], Gupta et al. [2, 3], Joshi et al. [7] also used the predecessor-successor technique and proposed different estimators in case of different probability sampling techniques for estimation of the characteristic under study without incorporating relative cost functions.

Here, the authors have tried to develop the estimator to derive the estimates of population total and its standard error in the case of simple random sampling with replacement (SRSWR) and without replacement (SRSWOR) schemes. Also, suitable

linear cost function is used to obtain the optimum values of the sample sizes, which have not been taken into consideration so far. The theory is hence supported with the help of numerical illustration.

2. Statement of the Problem

Let us consider a field survey, for enumerating the total wheat yield in a specific district. A sampling frame is available to the surveyor from which a sample has been selected for enumeration. But as soon as the investigator visits the sampled units, it is observed that there are some new agricultural lands growing wheat which were not reported in the original sampling frame. Thus it leads to the problem of incomplete sampling frame and the data based on new inclusions will affect the total crop estimation. Thus it becomes vital, to take into account, the new information for generating improved estimate of total wheat production.

In countries like India, the problem of incomplete frames is very common due to lack of technology and inadequate sources of data management and recording. Yates [14], in first edition of his book described incompleteness as a principle weakness of sampling frames. Hansen et. al. [4] gave the technique of predecessor-successor method deals with such problems. Despite the fact that incomplete sampling frames have been considered as the most important reason of imperfection in sample surveys, as also mentioned by numerous authors as Yates [14], Hansen et al. [4], Hartley [6] and many others, it was not taken into account on a larger scale. After Singh [13], the problem was not tackled till Agarwal et al. [1]. The problem still remains neglected to a very large extent. This study is an attempt to improvise the work and the estimators used so far in purview of cost constraints and efficiency.

Let Y be the characteristic under study, and let us consider a situation where the population consists of N_1 units listed in the frame. A sample of n_1 units is taken from the population from the existing frame. As mentioned above, during investigation it is observed that some new units have also been added to the population since the existing frame was constructed, say n_2. Let us assume that a total of N_2 (unknown) units is actually not included in the existing frame so that the actual size of the population is $N_1 + N_2 = N$ rather than N_1.

Now, using the predecessor-successor method, it is assumed that the N_2 units that are not included in the sampling frame occur somewhere in the gaps between the N_1 units in the population. Let U_i denotes the i^{th} unit included in the frame and $M_i, \forall i = 1,2,\dots,N_1$ denote the number of non-included units between the i^{th} and $(i+1)^{th}$ included unit, so that an arrangement as below is obtained

U_1 ——M_1; U_2——M_2;...;U_{N_1}——M_{N_1}.

From the definition of M_i, we have $\sum_{i=1}^{N_1} M_i = N_2; 0 \leq M_i \leq N_2$.

Investigation under such situation will definitely increase the cost for enumeration while including the non-included units (in existing sampling frame) in the study. In most of the studies, the cost of survey is fixed. Therefore, it is not possible to enumerate the entire lot of

new units without affecting the total cost of the survey.In the current situation, we take a sub-sample from the new units listed for possible inclusion and enumerate them. The cost for enumerating those units is balanced by reducing the actual sample size from the existing frame and utilising the same, to collect information from the non-included units.It is thus, requiredto optimise the two sample sizes (n_1 and n_2') so that new information can be inculcated in the study without affecting the cost of the survey. Hence, a suitable cost function is developed and optimum sample sizes by minimising variance for a fixed cost are determined.

Thus, the problem here arises that of

1. Estimation of the total number of non-included units in the population.
2. Determination of optimum sample sizes for a given cost.
3. Estimation of the population total of character under study.
4. Estimation of the standard error of the population total.

3. Estimation under SRSWR

3.1 Case 1: When the non-included units behave as included units

In this case, it can be assumed that the average of the characteristic for the non-included units is same as that of the included units. Here, the problem of incompleteness is taken into consideration for SRSWR scheme which was not originally taken into consideration by Singh(1983)(1989)or any other author in literature. For doing so, the estimator for mean given by Singh (1983)is used defined as,

$$\bar{y}_w = (1 + \bar{m})\bar{y}_{n_1} \tag{1}$$

where, $\bar{m} = \frac{1}{n_1}\sum_{i=1}^{n_1} M_i$ gives the average number of missing observations between each i^{th} and $(i+1)^{th}$ included unit. On multiplying N_1 and $\bar{m}$ the estimate of total number of non-included items in the population, $\widehat{N}_2$, is obtained as

$$\widehat{N}_2 = N_1\bar{m} \tag{2}$$

To estimate population total using (1) the statistic is defined as

$$T_1 = N_1\bar{y}_w. \tag{3}$$

The estimate of population total can then be obtained as,

$$\begin{aligned} E(T_1) &= E[N_1\bar{y}_w] \\ &= E[N_1(1+\bar{m})\bar{y}_{n_1}] \\ &= N_1\bar{Y}_{N_1} + N_2\bar{Y}_{N_1} \\ &= N\bar{Y}_N \because \bar{Y}_{N_1} \approx \bar{Y}_{N_2} \approx \bar{Y}_N \\ &= Y_N \end{aligned} \tag{4}$$

Thus it is seen that $T_1 = N_1\bar{y}_w$ is an unbiased estimate of population total Y_N. Now the variance of the estimateT_1 in case of SRSWR can be obtained as,

$$V(T_1)_{WR} = V(N_1\bar{y}_w)_{WR} = {N_1}^2 V(\bar{y}_w)_{WR} \tag{5}$$

to find the $V(T_1)_{WR}$, $V(\bar{y}_w)_{WR}$ in case of SRSWR can be obtained as follows:

$$V(\bar{y}_w)_{WR} = V[(1+\bar{m})\bar{y}_{n_1}] = V(\bar{y}_{n_1}) + V(\bar{m}\bar{y}_{n_1}) + 2Cov(\bar{y}_{n_1}, \bar{m}\bar{y}_{n_1}) \tag{6}$$

As SRSWR is considered,

$$V(\bar{y}_{n_1}) = \left(\frac{N_1-1}{N_1 n_1}\right){S_1}^2 \tag{7}$$

where ${S_1}^2 = \frac{\sum_{i=1}^{N_1}(Y_i - \bar{Y}_{N_1})^2}{(N_1-1)}$. Now $V(\bar{m}\bar{y}_{n_1})$ can be obtained as

$$V(\bar{m}\bar{y}_{n_1}) = V(\bar{m})\,V(\bar{y}_{n_1}) + V(\bar{m})\,[E(\bar{y}_{n_1})]^2 + V(\bar{y}_{n_1})\,[E(\bar{m})]^2 \tag{8}$$

where

$$V(\bar{m}) = \left(\frac{N_1-1}{N_1 n_1}\right){S_m}^2 \tag{9}$$

with ${S_m}^2 = \frac{\sum_{i=1}^{N_1}(M_i - \bar{M})^2}{(N_1-1)}$. Lastly, $Cov(\bar{y}_{n_1}, \bar{m}\bar{y}_{n_1})$ is given as

$$Cov(\bar{y}_{n_1}, \bar{m}\bar{y}_{n_1}) = \bar{M}V(\bar{y}_{n_1}) \tag{10}$$

Substituting (7), (8), (9) and (10) in (6) and using in (5),

$$V(T_1)_{WR} = N_1\left(\frac{N_1-1}{n_1}\right)\left[(1+\bar{M})^2{S_1}^2 + {\bar{Y}_{N_1}}^2{S_m}^2 + \left(\frac{N_1-1}{N_1 n_1}\right){S_1}^2{S_m}^2\right] \tag{11}$$

An estimate of the variance in this case is obtained as

$$\hat{V}(T_1)_{WR} = N_1\left(\frac{N_1-1}{n_1}\right)\left[(1+\bar{m})^2{s_1}^2 + {\bar{y}_{n_1}}^2{s_m}^2 + \left(\frac{N_1-1}{N_1 n_1}\right){s_1}^2{s_m}^2\right] \tag{12}$$

3.2 Case 2: When the non-included units do not behave as included units

In this case, let us suppose that n_2 units are found appearing amongst the gaps between the selected n_1 units in the sample, i.e., $n_2 = \sum_{i=1}^{n_1} M_i$. Then, a sample of ${n_2}'$ units is taken from the non-included n_2 units and obtain the data. Let $f = \frac{n_2}{{n_2}'}, such\ that (f \geq 1)$, be called as the retainment factor. Then let us define a new estimator T_2 as

$$T_2 = N_1\bar{y}_w{}'$$

$$\bar{y}_w{}' = \bar{y}_{n_1} + \bar{m}\bar{y}_{{n_2}'} \tag{13}$$

Also, it can easily be shown that, $E(T_2) = Y_N$, i.e. T_2 gives an unbiased estimate of population total.

The variance of the estimator $\bar{y}_w{}'$ in case of SRSWR can be obtained as,

$$V(\bar{y}_w{}')_{WR} = V(\bar{y}_{n_1} + \bar{m}\bar{y}_{n_2}{}')$$
$$= V(\bar{y}_{n_1}) + V(\bar{m})\,V(\bar{y}_{n_2}{}' + [E(\bar{m})]^2 V(\bar{y}_{n_2}{}') + [E(\bar{y}_{n_2}{}')]^2 V(\bar{m})$$
$$= \left(\frac{N_1-1}{N_1 n_1}\right)\left[{S_1}^2 + {\bar{y}_{n_2}}^2 {S_m}^2\right] + \left(\frac{n_2-1}{n_2{}' n_2}\right)\left[\bar{M}^2 + \left(\frac{N_1-1}{N_1 n_1}\right){S_m}^2\right]{S_2}^2 \quad (14)$$

An estimate of the variance can be obtained as

$$\hat{V}(T_2)_{WR} = {N_1}^2\left[\left(\frac{N_1-1}{N_1 n_1}\right)\left[{s_1}^2 + {\bar{y}_{n_2}{}'}^2 {s_m}^2\right] + \left(\frac{n_2-1}{n_2{}' n_2}\right)\left[\bar{m}^2 + \left(\frac{N_1-1}{N_1 n_1}\right){s_m}^2\right]{s_2}^2\right] \quad (15)$$

4. Estimation under SRSWOR

4.1 Case 1: When the non-included units behave as included units

Same as in case of SRSWR, it can be assumed that the average of the characteristic for the non-included units is same as that of the included units. Hence, statistic defined in (1) is used. Also, same as in SRSWR, it can be shown that T_1 gives an unbiased estimate of the population total Y_N and $N_1\bar{m}$ gives an estimate of the size of the non-included population, i.e. N_2. The variance of the estimator $\bar{y}_w$ is obtained as

$$V(\bar{y}_w)_{WOR} = V[(1+\bar{m})\bar{y}_{n_1}] = V(\bar{y}_{n_1}) + V(\bar{m}\bar{y}_{n_1}) + 2Cov(\bar{y}_{n_1}, \bar{m}\bar{y}_{n_1}) \quad (16)$$

$$\text{Here, } V(\bar{y}_{n_1}) = \left(\frac{1}{n_1} - \frac{1}{N_1}\right){S_1}^2 \quad (17)$$

$$V(\bar{m}) = \left(\frac{1}{n_1} - \frac{1}{N_1}\right){S_m}^2 \quad (18)$$

$$V(\bar{m}\bar{y}_{n_1}) = V(\bar{m})\,V(\bar{y}_{n_1}) + V(\bar{m})\,[E(\bar{y}_{n_1})]^2 + V(\bar{y}_{n_1})\,[E(\bar{m})]^2 \quad (19)$$

$$Cov(\bar{y}_{n_1}, \bar{m}\bar{y}_{n_1}) = \bar{M}V(\bar{y}_{n_1}) \quad (20)$$

Substituting (17), (18), (19) and (20) in (16) and solving,

$$V(T_1)_{WOR} = {N_1}^2\left(\frac{1}{n_1} - \frac{1}{N_1}\right)\left[(1+\bar{M})^2{S_1}^2 + {\bar{Y}_{N_1}}^2{S_m}^2 + \left(\frac{1}{n_1} - \frac{1}{N_1}\right){S_1}^2{S_m}^2\right] \quad (21)$$

An estimate of the variance in this case is obtained as

$$\hat{V}(T_1)_{WOR} = {N_1}^2\left(\frac{1}{n_1} - \frac{1}{N_1}\right)\left[(1+\bar{m})^2{s_1}^2 + {\bar{y}_{n_1}}^2{s_m}^2 + \left(\frac{1}{n_1} - \frac{1}{N_1}\right){s_1}^2{s_m}^2\right] \quad (22)$$

Here the expression for variance obtained for SRSWOR differs from the ones obtained by Singh (1983) which was

$$\hat{V}(T_1)_{WOR} = \frac{N_1(N_1-n_1)}{n_1}\left[\bar{y}_1{}^2{s_m}^2 + (1+\bar{m})^2{s_1}^2 - 2\left(\frac{1}{n_1} - \frac{1}{N_1}\right){s_1}^2{s_m}^2\right] \quad (23)$$

4.2 Case 2: When the non-included units do not behave as included units

Same as done in case of SRSWR, it is assumed that there exist n_2 units appearing in the n_1 gaps, out of which a sample of $n_2{}'$ units is taken for refining the results further. Thus, using the same statistic $\bar{y}_w{}'$ as defined in (13).Also, same as in SRSWR, it can be shown that T_2 gives an unbiased estimate of the population total Y_N.

Also, the variance of the estimator $\bar{y}_w{}'$ is obtained as

$$V(\bar{y}_w{}')_{WOR} = V(\bar{y}_{n_1} + \bar{m}\bar{y}_{n_2}{}')$$

$$= V(\bar{y}_{n_1}) + V(\bar{m})\,V(\bar{y}_{n_2}{}' + [E(\bar{m})]^2 V(\bar{y}_{n_2}{}') + [E(\bar{y}_{n_2}{}')]^2 V(\bar{m})$$

$$= \left(\frac{1}{n_1} - \frac{1}{N_1}\right)\left[{S_1}^2 + {\bar{y}_{n_2}}^2 {S_m}^2\right] + \left(\frac{1}{n_2{}'} - \frac{1}{n_2}\right)\left[\bar{M}^2 + \left(\frac{1}{n_1} - \frac{1}{N_1}\right){S_m}^2\right]{S_2}^2 \quad (24)$$

An estimate of the variance can be obtained as

$$\hat{V}(T_2)_{WOR} = {N_1}^2\left[\left(\frac{1}{n_1} - \frac{1}{N_1}\right)\left[{s_1}^2 + {\bar{y}_{n_2}{}'}^2 {s_m}^2\right] + \left(\frac{1}{n_2{}'} - \frac{1}{n_2}\right)\left[\bar{m}^2 + \left(\frac{1}{n_1} - \frac{1}{N_1}\right){s_m}^2\right]{s_2}^2\right] \quad (25)$$

5. Determination of Optimum Sample Size

The total cost of survey can be broadly categorised as

- C_0: The total establishment cost
- C_1: Cost per unit for n_1 included units
- C_2: Cost per unit for $n_2{}'$ non-inlcuded units

Thus the linear cost function is defined as

$$C' = C_0 + C_1 n_1 + C_2 n_2{}' \quad (26)$$

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

$$E(C') = C \quad (27)$$

$$E(C') = E\left(C_0 + C_1 n_1 + C_2 \frac{n_2}{f}\right)$$

$$C = C_0 + C_1 n_1 + C_2 \frac{n_1 N_2}{N_1 f} \quad (28)$$

where, $E(n_1) = n_1$ and $E(n_2) = \frac{n_1}{N_1}.N_2$

Here, the optimum values of n_1 and f are obtained for which variance is minimum for fixed cost C.

Considering, Lagrange's method for minimising variance as-

$$\phi = V(T_1)_{WOR} + \lambda[E(C') - C] \quad (29)$$

where, C is the given cost and λ is Lagrange's multiplier. To obtain the optimum values of $n_1 and f$ using the concept of maxima/minima and solving $\frac{\partial \phi}{\partial n_1} = 0$,

$$-\frac{1}{{n_1}^2}\left[{S_1}^2 + {\bar{y}_{n_2}}^2 {S_m}^2 + \frac{(f-1)}{n_2}{S_m}^2 {S_2}^2\right] + \lambda\left[C_1 + \frac{N_2 C_2}{N_1 f}\right] = 0$$

$$\lambda = \frac{\left[{S_1}^2 + {\bar{y}_{n_2}}^2 {S_m}^2 + (f-1)n_2 {S_m}^2 {S_2}^2\right]}{{n_1}^2[C_1 + N_2 C_2 N_1 f]} \quad (30)$$

Solving $\frac{\partial \phi}{\partial \lambda} = 0$,

$$C_0 + n_1\left\{C_1 + \frac{N_2 C_2}{N_1 f}\right\} - C = 0$$

$$n_{1_{opt}}(WOR) = \frac{C - C_0}{C_1 + N_2 C_2 N_1 f} \tag{31}$$

Solving $\frac{\partial \phi}{\partial f} = 0$,

$$\frac{1}{n_2}\left\{\bar{M}^2 + \left(\frac{1}{n_1} - \frac{1}{N_1}\right){S_m}^2\right\}{S_2}^2 - \lambda\frac{n_1 N_2 C_2}{N_1 f^2} = 0$$

$$\left[(C - C_0)(N_1\bar{M}^2 - {S_m}^2) + N_1 C_1 {S_m}^2\right]{S_2}^2 f^2 - \left[n_2\left\{{S_1}^2 + {\bar{y}_{n_2}}^2 {S_m}^2\right\} - N_2 C_2 {S_m}^2 {S_2}^2\right] = 0$$

$$f_{opt}(WOR) = \sqrt{\frac{n_2\left\{{S_1}^2 + {\bar{y}_{n_2}}^2 {S_m}^2\right\} - N_2 C_2 {S_m}^2 {S_2}^2}{\left\{(C - C_0)(N_1\bar{M}^2 - {S_m}^2) + N_1 C_1 {S_m}^2\right\}{S_2}^2}} \tag{32}$$

Here, (32) is solved to obtain f, subject to the condition that $f \geq 1$. Equations (31) and (32) provides the optimum values of n_1 and f under SRSWOR sampling scheme.

Similar to the optimisation of n_1 and f under SRSWOR, we find the optimum values under SRSWR sampling scheme, for which the Lagrange's function is defined as

$$\phi = V(T_1)_{WR} + \lambda[E(C') - C] \tag{33}$$

Here, assuming that N_1 and n_2 are sufficiently large so as to give us $\frac{N_1 - 1}{N_1} \approx 1$ and $\frac{n_2 - 1}{n_2} \approx 1$, we solve (33) to obtain the optimum n_1 and f and thus get

$$n_{1_{opt}}(WR) = \frac{C - C_0}{C_1 + N_2 C_2 N_1 f} \tag{34}$$

$$f_{opt}(WR) = \sqrt{\frac{n_2\left\{{S_1}^2 + {\bar{y}_{n_2}}^2 {S_m}^2\right\} N_2 C_2}{\left\{(C - C_0)\bar{M}^2 + C_1 {S_m}^2\right\} N_1 {S_2}^2}} \tag{35}$$

Here f_{opt} is that root of f, for which $f \geq 1$.

6. Numerical Illustration

This paper focuses on estimating parameters in the presence of incomplete sampling frames when sampling is done using simple random sampling techniques (with replacement and without replacement). The data Singh and Choudhary [11] in Table 1 shows the one complete lactation of milk yield (in 10 kg) of 250 cows in an organised dairy farm. Methods developed here are mainly concerned with estimating the population total as the estimator of the population mean is biased. To justify the applicability of the proposed estimation method following illustration is provided.

Table 1: One complete lactation of milk yield (in 10 kg) of 250 cows.

265	197	188	189	175	211	207	203	288	279	211	195	312	145	155	214	192	230	151
204	215	219	207	287	212	111	176	198	178	187	158	143	201	230	231	247	353	158
300	213	171	792	277	284	272	233	241	275	184	156	267	313	287	163	247	170	253
195	290	99	327	230	214	192	239	219	260	189	185	298	230	329	335	297	159	179
239	146	329	201	258	283	127	176	167	191	305	170	196	185	265	198	178	236	263
173	232	199	300	137	139	303	218	193	174	221	271	139	166	150	270	240	181	158
237	305	180	206	174	223	221	193	234	235	253	160	205	147	286	187	290	230	250
282	184	225	199	301	212	187	243	179	338	225	188	298	223	271	174	234	240	226
221	149	257	299	260	207	309	236	126	242	327	165	238	271	268	163	170	212	246
218	267	202	153	282	224	263	275	173	238	203	218	217	133	198	201	227	242	301

Data on incomplete/imperfect sampling frames are rarely available. Therefore authors have considered the above data (population size 250) to justify the results. This population is divided into two sub-populations, viz, included (N_1) and not included (N_2) in the sampling frame based upon Inc_{prop} (included proportion), which ranges from 0.5 to 0.9 where $Inc_{prop} = 0.9$ shows that 90% of the observations are included in the sampling frame. Here, a simple random sampling technique(WR and WOR) has been used for sampling units from the population. Sampling has been done based upon Sam_{prop} (sample proportion), i.e. the proportion of units sampled from included sub-population, which ranges from 0.1 to 0.3. If $Inc_{prop} = 0.5$ and $Sam_{prop} = 0.1$ then $n_1 = 13$(rounded) which is the smallest sample size studied in this paper. Enumeration of all possible samples in the case of with replacement which is 125^{13} and in the case without replacement which is $\binom{125}{13}$ is not possible here. Theoretically, we have shown that the sample total is an unbiased estimator of the population total. However, this can not be shown numerically here as enumeration of all possible samples is far beyond our reach in this paper. To resolve this issue, Monte Carlo simulations have been done.

Using the values of Inc_{prop} and Sam_{prop} Table 2 and Table 3 have been constructed which shows the average values of $n_{1_{opt}}$, $\hat{n}_2$ (estimate of non-included sample), $\widehat{N}_2$ (estimate of non-included population), $\hat{Y}_N^{N_1}$ (estimate of population total based on included population only), $\hat{v}(\hat{Y}_N^{N_1})$ (estimated variance of $\hat{Y}_N^{N_1}$), $\hat{Y}_N^{w}$ (estimate of population total based on included and non-included population), $\hat{v}(\hat{Y}_N^{w})$ (estimated variance of $\hat{Y}_N^{w}$) and efficiency for 50000 replications when sampling is done WR and WOR respectively. The population total of the data shown in Table 1 is 55906, and the variance of the population total is 255429664. As it can be seen in Table 1, for a specific Inc_{prop} variance of both the estimators decreases as Sam_{prop} increases. However, if we

look at the estimators, $\hat{Y}_N^{N_1}$ has been far deviated from the actual value in comparison to the $\hat{Y}_N^{w}$. As Inc_{prop} decreases the deviation of $\hat{Y}_N^{N_1}$ from the true value increases. On the other hand, $\hat{Y}_N^{w}$ is quite consistent. $\hat{N}_2$ is found equal to N_2 except when $Inc_{prop} = 0.9$ and $Sam_{prop} = 0.1$, which possibly could have happened because n_2 is small for $n_1 = 23$. Also, if we look at the efficiency, for all the combination of Inc_{prop} and Sam_{prop}, it is greater than one; hence our proposed estimator is performing far better than the existing estimator except when $Inc_{prop} = 0.5$ and $Sam_{prop} = 0.1$. This suggests that it is better to prepare a new sampling frame if more than 50% observations are found not included. Similar interpretations can be made for without replacement cases using Table 3. Comparing the two tables shows that the proposed technique for fixed cost is relatively more efficient for the SRSWOR case than in SRSWR.

Table 2: Average values of $n_{1_{opt}}$, $\hat{n}_2$, $\hat{N}_2$, $\hat{Y}_N^{N_1}$, $\hat{v}(\hat{Y}_N^{N_1})$, $\hat{Y}_N^{w}$, $\hat{v}(\hat{Y}_N^{w})$ and efficiency for 50000 replications when sampling is done with replacement.

Inc_{pro}	N_1	N_2	Sam_{pro}	$n_{1_{opt}}$	$\hat{n}_2$	$\hat{N}_2$	$\hat{Y}_N^{N_1}$	$\hat{v}(\hat{Y}_N^{N_1})$	$\hat{Y}_N^{w}$	$\hat{v}(\hat{Y}_N^{w})$	Eff.
0.9	225	25	0.1	22	2	23	56414	850602087	55397	22860590	11.86
			0.2	46	5	25	47426	134384972	55849	11869140	19.23
			0.3	70	8	25	48812	172308531	55894	7936359	27.98
0.8	200	50	0.1	19	5	50	42758	108259280	55903	40154308	4.34
			0.2	39	10	50	45728	81410230	55908	20349049	7.38
			0.3	59	15	50	42359	91721735	55922	13633747	10.35
0.7	175	75	0.1	19	8	75	41834	100387053	55946	53157015	2.38
			0.2	36	15	75	38728	76378443	55933	29275810	3.76
			0.3	54	22	75	39586	98342977	55923	19761676	5.25
0.6	150	100	0.1	17	10	100	30274	57947388	55962	72714783	1.29
			0.2	34	20	100	31857	64927078	55890	37857113	2.11
			0.3	51	30	100	32065	57435519	55934	25801060	2.93
0.5	125	125	0.1	15	12	125	25433	55685139	55907	105630908	0.71
			0.2	32	25	125	27789	46724060	55886	52061139	1.16
			0.3	49	38	125	27362	40204238	55890	34477347	1.63

Table 3: Average values of $n_{1_{opt}}$, $\hat{n}_2$, $\hat{N}_2$, $\hat{Y}_N^{N_1}$, $\hat{v}(\hat{Y}_N^{N_1})$, $\hat{Y}_N^w$, $\hat{v}(\hat{Y}_N^w)$ and efficiency for 50000 replications when sampling is done without replacement.

Inc_{prop}	N_1	N_2	Sam_{prop}	$n_{1_{opt}}$	$\hat{n}_2$	$\hat{N}_2$	$\hat{Y}_N^{N_1}$	$\hat{v}(\hat{Y}_N^{N_1})$	$\hat{Y}_N^w$	$\hat{v}(\hat{Y}_N^w)$	Eff.
0.9	225	25	0.1	22	2	23	47416	131823898	55435	20989282	13.29
			0.2	46	5	25	49921	122822120	55871	9885440	23.53
			0.3	70	8	25	48243	142726538	55905	5690416	39.60
0.8	200	50	0.1	19	5	50	44895	106152749	55933	36133248	4.65
			0.2	39	10	50	42431	110055870	55927	15653845	8.99
			0.3	59	15	50	41797	105255850	55911	9068966	14.53
0.7	175	75	0.1	19	8	75	36326	54734755	55960	49137054	2.55
			0.2	36	15	75	37674	76794569	55918	23044464	4.61
			0.3	54	22	75	39699	86668065	55904	13528161	7.39
0.6	150	100	0.1	17	10	100	30759	67467574	55939	68442122	1.39
			0.2	34	20	100	32881	76861063	55909	30177264	2.62
			0.3	51	30	100	35215	63569329	55915	17463493	4.25
0.5	125	125	0.1	15	12	125	29208	58972470	55975	102109965	0.76
			0.2	32	25	125	26035	38718986	55860	40886780	1.47
			0.3	49	38	125	27237	49151201	55928	22614831	2.45

7. Conclusion

As already shown above, the authors have successfully obtained refined and more efficient estimates of population total and their standard error for simple random sampling technique. The results obtained hereby help us conclude that the estimator used in the paper can readily be used in real life sampling problems where the problem of incomplete sampling frame arises. Also the same estimators can be used for an older data. Monte Carlo simulation technique provides us a clear picture of the problem and how it can be dealt in real life. It can here be concluded from the simluated results that a new sampling frame should be constructed if more than 50% observations are not included in the existing sampling frame.

Acknowledgement

The authors are thankful to the Referees for suggstions.

We would also 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., & Gupta, P. C. (2008). Estimation from Incomplete Sampling Frames in Case of Simple Random Sampling. *Model Assisted Statistics and Applications*.

[2] Gupta, P. C., Joshi, L., Joshi, V., & Nagar, P. (2019). Weighted Product Estimator for Incomplete Sampling Frame. *Journal of Rajasthan Academy of Physical Sciences*, 233-240.

[3] Gupta, P. C., Joshi, V., Nagar, P., & Singh, A. K. (2021). Linear Regression Estimator in Case of Incomplete Sampling Frame. *Journal of Rajasthan Academy of Physical Sciences*, 49-56.

[4] Hansen, M. H., Hurwitz, W. N., & Jabine, T. B. (1963). The Use of Imperfect Lists for Probability Sampling. *U.S. Bureau of Census* (pp. 497-517). Canada: Int. Stat. Inst. Ottawa.

[5] Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley and Sons.

[6] Hartley, H. O. (1962). Multiple Frame Surveys. *Social Statistics Section America Statistical Association*, (pp. 203-206). Minnesota.

[7] Joshi, V., Nagar, P., Singh, A. K., & Gupta, P. C. (2021). Use of Ratio Method of Estimation in Incomplete Frames. *International Journal of Agricultural and Statistical Sciences*, 267-272.

[8] Kish, L. (1965). *Survey Sampling.* John Wiley and Sons Inc., New York.

[9] Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of thw Royal Statistical Society*, 558-625.

[10] Seal, K. C. (1962). Use of Out-Dated Frames in Large Scale Sample Surveys. *Calcutta Statistical Association Bulletin*, 68-84.

[11] Singh, D., & Choudhary, F. S. (1986). *Theory and Analysis of Sample Survey Designs.* New York: Wiley.

[12] Singh, R. (1983). On the Use of Incomplete Frame in Sample Survey. *Biometrical Journal*, 545-549.

[13] Singh, R. (1989). Method of Estimation for Sampling from Incomplete Frames. *Australian Journal of Statistics*, 269-276.

[14] Yates, F. (1948). *Sampling Methods for Census and Surveys* (First Edition ed.). London: Charles Griffin and Co.