JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

REVIEW ON DEALING WITH THE PROBLEM OF IMPERFECT FRAME IN SAMPLE SURVEY

Singh Neelam Kumar, *Associate Professor, Department of Agricultural Statistics, BNPG College, Rath, Hamirpur, Uttar Pradesh, India.*
Email: neelamkumar099@gmail.com

Abstract: An appraisal of problem of imperfect frame in sample survey has been made. The causes and types of imperfection of frame have been discussed covering various aspects. The review of work done by different scholars has been explained under different situation of imperfection of frame. The sampling methodology under different situations of imperfection is explained and devised. Different measures and suggestions have been given to deal with imperfection of frame and estimators devised have been elucidated.

Keywords: Imperfect frame, Sampling* error*, Non-sampling error, Sampling frame, Sampled population, Target population

Introduction: Availability of sampling frame is a basic requirement for application of probability sampling technique so that each and every element in the population has a known and non-zero chance of being included in the sample. An ideal frame should consist of all elements occurring only once and should exclude any other element that is irrelevant for study. It is difficult to have a perfect frame in every situation. United Nations Secretariat (Statistical Division) presented draft report on sampling frames and master samples (Turner A. G., 2003) and explained the issues of sampling frame.

The results of the statistical analysis are always valid only for sampled population which corresponds to the list of sampling units, no matter whether conclusions on the sampled population are based on the sample survey or complete causes. The purpose of the statistics is however, to provide exact information on the target population. Therefore the population to be sampled (the sampled population) should coincide with the population about which information is wanted (the target population). To what extent, the conclusions drawn from the sampled population will also apply to the target population must depend upon the other sources of information. Hence, supplementary information about the nature of difference between sampled and target population must be gathered. Such information is available on the structure of the frame which is list of all the sampling units. Therefore, our first requisite is the existence of complete frame with reference to which relevant data are to be collected and represented consisting of description of all sampling units. The nature and details of the frame become the basis for the choice of appropriate sampling design. But frame, so constructed are often found to be incomplete, out-dated, illegible and contains unknown duplication. Bitter experience show that sampler have acquired a critical attitude towards the list that have been routinely collected for some purpose. Thus, there are rare situations in practice where frame in available in the form sampler desires to use. Incompleteness is common to virtually all lists used for sampling mainly due to dynamic nature of the population.

Therefore, Conclusions based on the sampled population which corresponds to the frame actually applied in census or survey amounts error and bias for the results of the target population, because of imperfection of the frame.

The precision of the statistical results has to be judged by total error, which as to the results to be obtained for target population consists of following components.

(i) Error due to deviation of target population from sampled population due to imperfection of the frame (ii) Sampling error and (iii) Non Sampling error. Errors caused due to component (ii) and (iii) have so far been largely discussed and analyzed in the literature. But error caused due to component has exceptionally been discussed and no much work has been done in this direction. It is in the light of component (i) that draws attention of present study rarely discussed so far.

Error (i) arises from the fact that the sampled population to which results refer and the target population for which results are needed do not conform to each other due to imperfection and incompleteness of the frame. This error can broadly be classified as deviation of coverage and deviation of content. Error of deviation of coverage may occur due to omission of sampling units from the target population, sampling units being out of scope or out dated due to dynamic nature of the frame and duplication of units. Error due to deviation in content may occur when available frame provides incorrect auxiliary information on reporting units. By the time of actual survey, there is seldom one to one correspondence between units of sampled population and target population because of the dynamic nature of the population. Selection of sample from such imperfect frame will not subscribe to the principle of random selection as some of the out dated units which should have been assigned zero probability of selection have also been selected in the sample and it was discovered because unit was selected in the sample. Quite often a frame consists of some superfluous units which do not exist in the target population at the time of actual enumeration and for which rule of association do not lead to any of the reporting unit in the target population. The listed units which are associated with some reporting unit may be identical in all the respects so that imperfection arises due to duplication of units in the frame. Error caused due to imperfection of frame is serious when exceedingly large unit is incorrectly assigned a small measure of size when selection is with probability proportional to measure of size and error is discovered because the unit was selected.

Thus, imperfect frames are the rule rather than the exceptions. However, sampling theory appears to have been evolved largely around perfect frame based on ideal conditions. But in practice, the frame available is often incomplete and imperfect and will not conform in its delimitations and compositions to the target population. It remains, never the less, the aim of the statistics to supply information as exact as possible on the target population.

Thus, there is hardly a situation in which frame is up-to-date and perfect. The result of survey is valid foe sampled population but the objective of the statistics is to provide exact information about target population. Conventional survey sampling methodology assumes at least conceptually the availability of complete frame,

The problem of incomplete frames and the consequences of the imperfection of the frame in simple survey have attracted the attention of several workers especially in the developed countries.

Mahalanobis (1944) also discussed about sampling frame.

Yates, F. (1948) gave an early account of the principal weaknesses of the frames, in his first edition of the book. He pointed out that frames are often incomplete, that they sometimes contain superfluous units not in the target population, that they may have unknown amount of duplication and that the supplementary information supplied for stratification or unequal probability selection may be in error. He gave in compact form some sound practical advices for handling the imperfection.

Goodman Leo. A. (1952) considered problem of duplication in the frame by dividing the population into number of mutually exclusive classes. A random sample of size n without replacement was drawn and problem was to estimate total number K of classes which subdivide the population into mutually exclusive classes. According to Good man there is exactly one real valued statistic S which is an unbiased estimate of K when sample size is not less than a specified number q (say) of elements contained in any class.

$$\text{Let } S' = N - \frac{N(n-1)}{n(n-1)} X_2 , \; S' = \sum_{i=1}^{n} x_i$$

$$T = \sum_{i=1}^{n} x_i , \qquad \text{if} \quad S' < \sum_{i=1}^{n} x_i$$

Where, x_i is the number of classes containing I elements in the sample, is the most suitable estimate in comparison with three other statistics, for a hypothetical population. In this way after knowing independent number of classes, duplication can be estimated.

Harsen, M.N., Hurwitz, W.N., Madow, W.G. (1953) described the sample that covers whole population in 1949 sample survey of retail stores taken by Census Bureau by an example of sample from a list B of large business in combination with a sample from an a real frame A that covers the complete population. The objectives, in this combined use of an incomplete list frame and complete areal frame were to gain increased accuracy and save money. In retail stores surveys, the business in the list frame that was present in the area sample, so that population being sampled fell in to two distinct strata. This process of induplication was performed by the field supervisor in advance of sampling.

Yates, F. (1953) also gave basic formula for the study of the subpopulation or domain of study as represented in all strata.

Durbin (1958) and Hartley (1959) further discussed and gave proofs for the formula. Durbin (1958) pointed out to the means estimated for the whole population, if the sample is incomplete for any reason such as non – response.

Edward Deming and Glasser, G.J. (1959) proposed a method of matching the lists when two or more lists are available for uses. They presented theory of estimation of proportion of names common to two lists of names, through use of sample drawn from lists. They gave probability, expected value, and variance of estimates of proportions of duplicated units. Besides this they gave optimum allocation of samples, effects of duplicates within a list and possible gain from stratification.

Suppose, there are two or more lists of names, some names may be common to some or all of the lists. It is of some economic or scientific importance to discover how many. For example, a firm may wish to determine by comparing two lists, how many of their present employees worked there in some past year, the number of shareholders common to two or more companies and number of companies that do business in each of two states, which requires the problem of matching lists. The list may be very large and in practice it may run several hundred, thousand of millions of the names. A publisher of a magazine of a magazine who wished to discover how many of his subscribers were on a list of executives and on other special list.

They presented some statistical theory for solution of such problems. The theory was based on probability sampling from both lists which included sample from one list matched against the other.

Let a_1, a_2 , …………………… a_n are district and ordered names on one list and b_1, b_2, …………… b_n are distinct and ordered names on the other list such that D names are common to both the lists. They divided the first and second list such that no name appears more than once on other list.

Let $P = \frac{D}{M}$ and $P = \frac{D}{N}$

Let names in the sample be x_1, x_2………………….. x_m from list 1 and y_1, y_2, ……. Y_n from list 2.

Σ x_i y_i = 1 if the two names are identical

= 0 otherwise

They defined d =Σ x_i y_j { I =1………..m, j =1………n }

As the number of names common to the 2 samples so that d is a random variable.

Sampling procedure elucidated as: Draw a sample of m names between 1 to M by S.R.S.W.O.R from list 1 and n names from list 2 between 1 to N by S.R.S.W.O.R. compare every name in the sample list 1 with every name in the sample from list 2to discover how many names are common to both samples. Let d be this number.

Then from the estimates

$$\hat{p} = \frac{N}{n}.\frac{d}{m}, \qquad \hat{p} = \frac{M.d}{n.m}, \quad \hat{D} = \frac{NM}{n.m} d = M\hat{p} = N\hat{p},$$

Then

E($\hat{p}$) =p, E($\hat{D}$) = D

The probability distribution of d, number of names common to 2 samples.

$$P(d) = \frac{\binom{D}{d}}{\binom{M}{m}\binom{N}{n}} \sum_{k=d}^{D} \binom{D-d}{k-d}\binom{M-D}{m-K}\binom{N-K}{n-d}$$

This can be used to determine Est $v\left(\hat{p}\right)$, Est $v\left(\hat{P}\right)$ and critical values for statistical test and to compute the power of the test.

Again let A_1, $A_{2,}$ ……….. A_n be the D names common to both the lists. Then the probability that any specified set of K names will fall into both samples is

P (A_1, $A_{2,}$ ……….. A_k) = $\dfrac{\binom{m}{K}\binom{n}{K}}{\binom{n}{K}\binom{N}{K}}$ for any set of K names,

$K \leq D$, $K \leq m$, $K \leq n$.

From this also it can be seen that probability distribution of exactly d names common to both samples is same as given earlier. This procedure can be used to determine the number of duplicate units in the two frames.

Seal (1962) discussed the use of out – dated frame in large scale surveys. In a dynamic population, sample drawn does not provide unbiased estimate of the population characteristic existing during the period of survey. He suggested simple method to deal such problems on the basis of a reasonable birth and death process for taking into account checking population which could be worked out in most of the practical situations. Thus he considered change in the population as a continuous stochastic – process and suggested use of successive frames to allow for the change in the population when estimating from incomplete frame.

In many practical situations, available frame is not up –to –date as the new units established subsequently to the reporting period in which the list is compiled, are not included in the frame. As the population changes with passage of time and chosen sample fails to take account of any new unit that might come up after preparation of the out – dated frame, the estimate of any characteristic of the population existing during the period of survey on the basis of the selected sample is likely to be somewhat biased. Another difficulty arising out of the use of such a sample is that a large per centage of sampled units might be found to be out – dated at the time of canvassing these units. The frame relating to later period will make it possible to find out birth and deaths which account during the intervening period. However, assuming some reasonable model, It might be possible to drive the number of birth and deaths of units that could be expected on the basis of data available for two consecutive frames. He discussed the problem of estimating the total of average of any characteristic of a dynamic population at a particular point of time on the basis of such model.

He formulated the problems as follows:-

Let a survey is launched at time k and completed at time m (m> k).

Let information on certain characteristic y_t at time t is collected for chosen reference period t ϵ (a, b) where, a<b≤ k. Suppose, the frame from which sample selected related to time 0 so that a≥0, Suppose, at the analysis stage of data collected during the survey, the latest frame available relates to time p (p≥0).

Suppose that the frame contained some measurable characteristics denoted by x_1, x_2, ……….. $x_{r.}$ Let N_o and N_p are assumed as number of units at time o and p so that N_{op} are the number of units common between N_o and N_p .

Hence, N_o - N_{op} = number of deaths and N_p and N_{op} = number of births of units during period 0 and p.

Let a sample of N_o units was taken at random from N_o at time O. Let n_O^* be the total number of units for which necessary information could be collected which would imply that $N_O - N_O^*$ units have gone out of existence during time [a, b]. Information collected from these n_O^* units during the reference period is utilized to estimate population characteristic existing at time t, t ϵ (a, b). Suppose that a characteristic y_t to be studied in a dynamic population at time t constitute a stochastic – process $\{y_t\}$ and problem is to estimate $E\{y_t\} = A_t$, say, t ϵ (a, b) for the stratum on the basis of n_O^* units surviving from an out – of – date frame containing at time t = O.

Let N_t = number of units existing at time t in the population and

$\bar{y}_t = \frac{y_t}{N_t}$, the value of characteristic per unit population, Assume that y_t and N_t are uncorrelated for all t, so that;

$$Y_t = \bar{Y}_t N_t \text{ or}$$

$$E(Y_t) = E(\bar{Y}_t)E(N_t)$$

Or $A_t = M_t n_t$, where, $M_t = E\ (\bar{Y}_t)$ and $M_t = E(N_t)$

Therefore,

$$\bar{A}_t = \bar{M}_t \bar{n}_t$$

For estimation of M_t it was assumed that changing population of units $\{N_t\}$ existing of different point of the time constitute a birth and death process. For such a stochastic – process it was shown that –

$$M_t = \frac{\lambda}{\mu}(1-e^{-\mu}t) + M_O e^{-\mu}t$$

And estimates as λ and μ where shown to be as

$$\bar{\lambda} = \frac{\bar{\mu}\bar{N}_O(N_p - N_{op})}{N_o - N_{op}} \qquad \text{and}$$

$$\bar{\mu} = \frac{1}{p}(\log_e N_o - \log_e N_{op})$$ and

$$E(n_o^*) = n_o\left[\frac{e^{-\mu k} - e^{-\mu m}}{\mu(m-k)}\right]$$

The population average $n_t = E\,(y_t/N_t)$ of the characteristic y_t under study is usually to be estimated from sample average based on n_O^* units *i.e.* surviving units for which required information could be collecting during course of survey.

But as this sample of n_O^* units fails to take into account any new units that might come up after the preparation of the frame from which the original sample of n_O units were drawn at time, t =0, the sample average worked out on the basis of n_O^* units existing during the period of survey may not be an unbiased estimate of n_t particularly when the characteristic under study for newly established units posses certain special features. For instance, size of newly established units may be either very large or small; the average value based on the n_O^* units may be quite different from the average characteristic revealed by these new units. But if n_O^* were stratified taking into account all the reasonable factors such as size and location of units which could influence the characteristic under study, then estimation of population average n_t on the basis of n_O^* surveying units may provide reasonable good results. In estimating n_t on the basis of n_O^* units, two distinct cases would arise. One, when characteristic y under study is not correlated to any other auxiliary variable x and two, when y is correlated to one or more of the available auxiliary variable. In first case, y_t will be estimated by n_t. In second case, he advised to use method of double sampling or sample ratio estimates for estimating n_t.

Hartley (1962) proposed the use of two or more frames to overcome the problem of incomplete frames such that the entire population is covered by the use of these frames. He discussed the optimum sample size to be taken from the two frames under a suitable cost function.

He considered the identification of the sample members from one frame A sample that belong to the frame B. Then sampler has three samples at his disposal in which y has been measured - a samples from stratum a (the part A that does not belong to B) and two independent samples from stratum B. ab was the sample obtained from A and identified as belong to B, and one is obtained by direct sampling of frame B. He proposed that both B samples be used in the post – stratified estimate,

$$\hat{Y} = N_a \bar{Y}_a + N_{ab}(p\bar{Y}_{ab} + q\bar{Y}_B)$$

Where,

$\bar{Y}_a$, $\bar{Y}_{ab}$, $\bar{Y}_B$ denote the respective sample means. If frame a be incomplete, two frames A and B with some duplication , then to obtain complete coverage of the population, the three strata cannot be sampled directly but samples of sizes n_A, n_B have to be drawn from A and B. He suggested the estimate

$$\bar{Y} = \frac{N_a}{n_a} y_a + p\frac{N_A}{n_a} y_{ab} + q\frac{N_B}{n_B} y_{ba} + \frac{N_B}{n_B} y_{ba} + \frac{N_B}{n_B} y_B$$

Where,

P + q = 1 and

a : units in A alone

ab : units in both A and B found in n_A

b : units in B alone

ba : units in the duplicate stratum found in n_B.

He determined p and q to determine v ($\hat{Y}$) for fixed cost.

Hansen, M.H.; Hurwitz, W.N. and Jabine, J.B. (1964) have discussed at length about various types of imperfections arising in the frame. He advised various procedures for the use of the incomplete frame for probability sampling at the U.S. Bureau of census and suggested different methods to complete the frame for different kinds of the incompleteness arising in the frame.

They explained that, even if list of sampling units *i.e.* frames are incomplete. They often have some desirable features as:

(1) Degree of clustering of sample units can often be fully controlled where as in case of area sampling this degree of control is not feasible.
(2) Lists frequently have measure of size which are more highly correlated with items to be estimated from the samples than are the measures of size available from segments or other small areas used as ultimate sampling units in an area sample.
(3) Address of reporting units included on a list data can be collected by mail or telephone which is not possible in area sampling.
(4) Selection cost is small in samples from lists.
(5) In sampling from items that are rare in general population. List can lead to significant reduction in the cost of achieving desired reliability.

They preferred use of list in sampling even though lists are seriously incomplete due to several reasons as: (i) Area sampling techniques are less efficient than the use of available lists, even though, the lists are frequently incomplete. (ii) Increased availability of lists providing access to target population. (iii) Development of more efficient techniques for the joint use of samples from lists in combination with area samples. They described some of the principles and methods which were being used in survey at the US Census bureau to take advantage of lists inspite of their imperfections in the design of probability samples.

They suggested that if lists whose units are not necessarily those in the target population *i.e.*if a list is used for the purpose other than those for which it was originally designed , some rule of association between listed units and those in the target population should be established. He considered a lists with units 1_1, 1_2 ……… 1_m and target population with reporting units t_1, t_2, …….. t_n. A rule of association should be established between 1_1, 1_2 ……… 1_m and t_1, t_2, …….. t_n in such a way that selected of listed unit 1_i with known probability leads directly to the selection of reporting units t_j with known probability.

Further, if some units are seen to have been missed and thus have zero probability of selection, steps have to be taken to reach these units by sampling from some other sources.

They considered the case in which list is containing units which are not in the target population and suggested to remove units known to be out-of-scope form the list prior to sampling. They also suggested to follow re-sampling periodically form an up-to-date list and to treat as separate strata groups known to contain high proportion of out- of –scope units. For list containing erroneous and incomplete entries, they advised to treat listed units with incomplete information as out-of-scope and rely on complementary sample to cover the associated reporting unit using area sample.

In case of duplication, they advised to take all possible steps to eliminate duplication before sample is selected. When extent of duplication in the list is unknown, it is possible to select a sample of moderate size and use this to eliminate the extent of duplication in the sample list. They advised that if either number of duplicate is selected in the sample, it can be identified on the basis of information obtained from the reporting unit and true probability of selection for each reporting unit in the sample can be determined to obtain in unbiased consistent estimate.

That a list does not contain or lead to all units in the survey for target population and if we purpose to use the lists for sampling to get at units in target population, this incompleteness becomes an imperfection. In such cases they advised to redefine the target population and survey population to include only those units which were associated according to the rule of association established at the time of survey. Thus they proposed solution by redefinition of survey population in terms of an available list or lists. It is essentially the same as "cut-off" sampling procedure with cut-off defined in terms of the available lists. They advised to supplement an incomplete list by selection of sample units from another list, by finding a way to induplicate the samples from the two lists. They also advised to supplement incomplete list with area samples or joint use of list-area sample to achieve full coverage of target population at minimum cost. They reported that the survey design must include a procedure for identifying those reporting units that are accessible through both the lists and area of samples and gave a concept of induplication. List and area samples are unduplicated by removing from the area samples any reporting units that could have been reached through the list sample, whether or not they were actually selected.

They said that there are advantages to be gained from over-lapping of the list and are samples. By over-lapping the list and area sample, they meant selecting them in such a way that the units in the list sample have a higher probability of being located in or near units of the area sample than would be the case if the two samples were selected entirely independently. There are two advantages to such overlap. First, it reduces unit costs. Secondly, it may reduce the risk of bias in the induplication process. Again these giants, one must balance the losses in sampling efficiently which occurs because the units in the list sample must be selected in clusters in order to achieve overlap with area sample.

They proposed that the higher degree of overlap can be achieved through the use of segment list sample technique. This technique consists of placing a system on a map to show approximate location of each unit on the list and then identifying those units whose samples fail in the segments of the area sample which is to be used in the same survey.

They proposed that units in the area sample which are also represented by the list sample, do not necessarily have to be discarded in making estimates. They considered a population of N units where N = N1 + N_2 + N_3, where N_1 being number of units associable through list sampling, N2 through area sampling and N_3 through list and area sampling both.

Let X be some characteristic of population with $X = X_1 + X_2 + X_3$ where subscripts have same meaning.

Let $X_L^{'} = X_{L_1}^{'} + X_{L_3}^{'}$ as unbiased estimate of $X_1 + X_3$ based on a list sample and $X_A^{'} = X_{A_2}^{'} + X_{A_3}^{'}$ an unbiased estimate of X_2 and X_3 based on an area sample. Than unbiased estimate of X can be formed as:

$$X^{'} = X_{L_1}^{'} + X_{A_2}^{'} + KX_{L_3}^{'} + (1-K)X_{A_3}^{'}$$

A determination of the optimum value of K will result, in its taking as value other than 0 or 1, so that it will be desirable to base the estimate of X_3 on the sample units in both the list and area survey. They proposed that in a survey taken on successive occasions, using list sample supplemented by a complementary area sample, the sampling error component resulting from complementary area sample can often be substantially reduced by the use of sample rotation.

A method to include those units in the sample which are not contained in the frame, called predecessor-successor method to obtain information on omission in the frame was also proposed by them. They considered a population of reporting units divided into two classes, those that area included on an available list and those that are not on this list. A principle was supposed to be established *i.e.* a geographic ordering of units in the population and further rule of ordering are such that given any one unit in the population, one can uniquely determine its successor by following a defined path of travel. The procedure consists in first selecting a random sample from the available frame and then for each selected units determine its successor and check to see if it is on the list or not. If the successor is on the list, discard it and if the successor is not on the list, include it in the sample and then identify its successor and proceed further in the same way until a successor is found to be on the list. Thus, the sample will consist of those units in the original sample from list plus all sequence or units not on the list which immediately follow these units on the path of travel. The probability of selection of any unlisted unit, therefore, is the same as that of the first listed unit immediately proceeding in the path of travel. But the size of sample becomes a random variable and work load is increased in this method.

They also advised to deal incompleteness of the frame by sampling from decentralized lists *i.e.* selection of samples from decentralized records such as buildings permits, assessments records, vehicle registration etc.

Szameitat and Schaffar (1963) discussed about imperfect frames and consequences their use in sampling. They discussed at length about different types of imperfections arising in the frame and different types of errors caused due to imperfect frames in the statistical results. They elucidated that errors due to imperfect frames lead to error due deviation of target population from sampled population. They classified error caused due to deviation of sampled population and target population as (1) deviation of coverage and (2) deviation of content. The possibilities for reducing such errors with description of different error components was also advised on the basis of samples selected from such frames. A sample error model was provided for the various types of the errors caused by imperfect frames followed by presentation of some general principles of practical work.

They devised theoretical error analysis by means of the error model. Accordingly let the parts of the target population will be given the number g = 1, 2, 3…….. , k and parts of sampled population, the numbers h = 1, 2, 3 ……L (K ≤L).For g = 0 implied the number for the set of units in the sampled population which are not contained in the target population. Finally h = 0 designated the set of units in the target population which are not contained in the sampled population.

Let, the groups formed by the cross classification of the units is indicated as N_{gh}^{*} such thatit denoted the number of sampling units contained in the hth part of sampled population and belonging to gth part of sampled population and belonging to gth part of the target population. Generally, $N_{oo}^{*} = o$. Let, N_{oh}^{*} = number of sampling units contained in the part of sampled population and not belonging to the target population, but wrongly being considered as part of it. Let $N_{(g)\,h}$ = number of sampling units which are contained in the hth stratum of the sampled population, do not belong to the target population but are wrongly allocated to the gth part of target population. It was presumed that $N_{(g)0}^{**} = o$, Let n_{gh} = number of units contained in the hth statum of the sampled population and belonging to the gth part of the target population which is also reffered as ghth group.

Hence,

$$N_{gh} = N_{gh}^{*} + N_{(g)h}^{**} \qquad (g = 1, 2, \ldots\ldots\ldots.., k, \qquad h = 0, 1, 2, \ \ldots\ldots\ldots\ldots L)$$

Sampling units contained in the gth part of the target population

$$N_q = \sum_{h=o}^{L} N_{gh} , \qquad N_g^{*} = \sum_{h=1}^{L} N_{gh}^{*} ,$$

$$N_g^{**} = \sum_{h=o}^{L} N_{(g)h}^{**} ,$$

If, X_{gh}^{*}, $X_{(g)h}^{**}$ are the total values of the investigated characteristics with subscripts having the same meaning, then total value X_{gh} of all units of the ghth group can be given as:-

$$X_{gh} = X^{*}_{gh} + X^{**}_{(g)h}.$$

The total values of these domains of study are

$$X^{*}_{g} = \sum_{h=o}^{L} X^{*}_{gh}$$

Hence, total value $X^{*} = \sum_{g=1}^{k} X^{*}_{g}$

is estimated for the target population on the basis of imperfect frame.

A formula for systematic error due to incompleteness of the frame was devised. These errors were conditioned because of two factors (i) by share of units missing in the frame in the gth part of the target population and (ii) by the relation between the mean value $\bar{X}_{go}$ for the units missing in the frame in the gth part of the target population and the mean value $\bar{X}^{*}_{g}$ of the characteristics concerned in the gth part of target population.

They described that random sampling error of a sample survey may be considerably increased by deviations of the frame in content and also by deviations in coverage.

The S.R.S W.O.R. was drawn separately from each of the L parts of the sampled population and using the sampling fraction f_n in the nth stratum, a simple unbiased estimator was used for estimating the total value.

Hence, estimator for $\bar{X}_{g.}$ was given as:-

$$X'_{g.} = \sum_{h=1}^{L} \frac{X_{gh}}{f_{h}} \text{ so that}$$

$$E\left(x'_{g}\right) = \bar{X}_{g.}$$

Here, small letters are given sample values. x_{gh} for instance is the total value of the characteristics for the units of the ghth group included in the sample.

It is therefore, (according to the formula)

$$E\left(x'_{g.}\right) = X^{*}_{g.} - X^{*}_{go} + X^{**}_{g.}$$

They also devised formula for error variance of $X'_{g.}$.

They also advised to follow some general principles and rules which arise from the use of incomplete frame in sample and other surveys for practical work. They advised to perform two tasks, *i.e.* description of the error components and reduction of the error components apart from the ascertainment of the result desired. They primarily emphasized and investigated about error component caused due to deviation of the target population from the sampled population. For reducing the deviation between target population and sampled population and sampled population, they advised that,

(i) Prior to the execution of the survey, the frame should be checked and corrected on the basis of further material in the best possible way.

(ii) The procedure of area sampling offers in particular the possibility to cover, under certain circumstances, all reporting units of the target populations to ensure that all units newly accruing in the target population. In some demographic studied of New Guinea, area sampling proved to be fruitful.

When target population changes in time, particularly when there is long interval between the date to which the sampled population relates and the date or period of time for which information is to be collected, the area sampling cannot ensure due consideration of changes occurring in the target population (due to ascertainment of fluctuations, changes of reporting units from one subpopulation to another). To avoid such errors they advised, (i) Supplementation of the frame by including material newly according units in the target population or (ii) Insertion of complete censuses at certain intervals in order to determine the new position of the target population and to obtain an up – to- date sampling frame for current sample survey.

Lund (1968), Fuller and Burmeister (1972) gave an improvement to the Hartley's estimate using better estimate of Na, Nab, Nb, that are implied in Hartley's estimate. Fuller and Burmeister also dealt with the case in which frame A is a real unit.

Sen. A. R. (1970) conducted a special sample survey of Wabefowl in Ontario during 1968-69 and designed to estimate the bias if any resulting from the use of 1966-67 list of permit holders as sampling frame for the estimation of characteristics at the annual harvest for 1967 in Ontario in the Canadian waterfowl survey. It was estimated that the national estimates of kill of ducks for Ontario for 1967-68 has an upward bias to the extent of 8.5 per cent.

Hartley (1974) gives a general approach to two – frame sampling, applicable to any sample design in the two frames.

Singh, Randhir (1983) gave a mathematical formulation of the problem predecessor – successor method for estimating, the total number of units of the target population missing from the frame, and the total for the character under study for the target population.

In large scale surveys generally two different situations regarding units of the target population missing from the frame may arise,

(i) When the units missing from the frame are random *i. e.* these do not differ significantly from the units available in the frame with respect to the character under study, and

(ii) When missing units differ significantly from the units present in the frame.

He has given separate estimators for dealing with these two situations.

Singh R. (1989) Suggested estimation procedure when sampling is done from an incomplete frame which does not contain some of the units of target population and contains some extra units which donot belong to the target population. He Proposed estimation by predecessor- Successor method assuming that geographical ordering of units can be established and rule of ordering are such that given any one unit in the population, we can identify its successor by following a defined path of travel.

Bandyopadhyay. S and Adhikar A. K. (1993) discussed about imperfect frame in which no population unit has been excluded from the frame but an unspecified number of population units might have been included in the list an unspecified number of times each with a separate identification. It was established that for estimation of a population ratio are mean, the mean square errors of estimators based on the imperfect frame are less than those based on perfect frame for simple random sampling when the sampling fractions of perfect and imperfect frames are the same. It was shown that, there are situations in which estimates of a ratio, a ,mean or a total based on smaller sampling fraction from imperfect frame can have smaller mean snare error than those based on a larger sampling fraction from the perfect frame.

Singh, N. K., et al (1997) discussed imperfection in the frame of finite population and proposed estimators for domain of study considering probability distribution of the out-dated units in the incomplete frame.

Singh, N. K. et al. (2001) discussed the imperfection of frame arising due to omission some of the units from the frame and also frame containing some units which no more belongs to target population and proposed appropriate estimation procedure for the population, its variance when sampling is done from two frames.

Terril. Byczkowski, Martin S. Levy (2009) discussed the issue of Imperfect sampling frames when it can result in more efficient estimators of population total than perfect frame is explored. It was shown, how and when it provides an illuminating basis for comparing a weighted estimator under an imperfect frame with that of a conventional estimator assuming the frame has been corrected. It explained the circumstances under which an imperfect frame results in estimates of population total that are more precise than those from a perfect frame. A classification tree methodology was used to explore circumstances under which imperfect frame results in more precise estimators. The results complement, strengthens and explain that lead to recommendations as to when to correct a frame or when to adjust for imperfection uses a weighting methodology called the arc weight estimator.

Mecatti, F. Singh. A. C. (2014) explained and illustrated multiple frame survey as a total for dealing with imperfect frame. It was established that multiple frame surreys are useful for reducing cost for given precision constraints, improving coverage (Under or over) and dealing with elusive or rare populations for which a direct sampling frame may not exist. They provided a simplified and unified review of different existing methods which should help in better understanding in choosing a suitable method in any application and promoting more use of multiple frames in practice.

Singh, N. K. (2020) discussed the frame error as error due to imperfection of the frame in detail because of deviation of the target population from sample population.

Singh, N. K. (2020) discussed that frames are often imperfect in any sample survey which arises due to some of the rare sampling units being out-dated at the time of actual survey. He further devised unbiased estimator for the imperfection of frame arising due to rare out-of-scope units considering population size to be large. Suitable estimators for the proportion of out-dated units from the population and target population total for a character with their variance was developed considering probability distribution function (p.d.f.) of rare out- dated units in large population.

Singh, N. K. (2021) discussed that the existence of the frame is pre-requisite for any sample survey or census of a large population. Frames are quite often imperfect due to dynamic nature of sampling units. Frames become incomplete by the time actual survey and enumeration starts which affect the statistical results desired for the target population. He reported and considered imperfection in the frame of large population arising due to the qualitative change of units from one class to other. He considered incomplete frame assuming the nature of units following dynamic change from class one to other which follows a probability distribution function. Suitable estimator for proportion of units belonging to a particular domain and unbiased estimate of target population for a class was proposed along with its estimate of variance. The estimates are evolved so as to eliminate error caused due to the deviation of sampled population from target population.

It is assumed that such rare units missing from the frame might have contributed much to the target population parameter if it were not out-of-scope or out-dated units. These rare units missing from the target population may be considerably of high measure in their size and weight. Even each of the rare units missing may be of large size. More troublesome are those cases when missing units although rare in number are exceedingly large in their measure of size and such rare units were discovered because units were selected. For example, in list of business establishments, some large establishments although rare in number, may no longer be active at the time of enumeration. These large sized rare units missing from the frame would lead to the imperfection of the frame and would contribute much to the bias of sampling results for target population parameters.

References

Bandyopadhyay, S. and A. K. Adhikari (1993). Sampling from Imperfect Frames with Unknown Amount of Duplication. *Survey Methodology, December 1993. Vol. 19, No. 2, pp. 193-197. Statistics Canada*.

Byczkowski Terri L. and Levy Martin S. (2009). When Does an Imperfect sampling Frame Produce More Efficient Estimators than a Perfect Frame? Journal of Statistical Planning and Inference, Volume 139, issue 10. 1- October 2009, pages 3679-3689.

Deming E. W. and Glasser G.J. (1959). On the problem of matching lists by samples. J.A.S.A. Vol. 54, pp- 403-415.

Durbin, J. (1958). Sampling Theory for estimates based on fewer individuals than the number selected. *Bull. Int. Statistics Inst., 36, pp- 113-119.*

Fuller W.A. Burmeirster L.F (1972). Estimators for samples selected from two overlapping frames. *Proc. Soc. Stat. Sect. Amer. Stat. Associated, pp- 245-249.*

Fuvia Mecatti and Avinash C. Singh (2014). Estimation in Multiple Frame Surveys: a Simplified and Unified Review Using the Multiplicity *Approach. Journol De La Societe Francaise De Statistque. Vol. 155, No. 4 (2014).*

Leo A. Goodman (1952). On the Analysis of Samples from k Lists. Ann. Math Statist, Vol. 23, pp- 632634.

Lund R. E. (1968). Estimator in multiple frame survey. *Proc. Soct. Sci. Amer. Stat Assoc. pp- 282-288*.

Hansen M.H., Hurwitz W.N. and Jabine J.B. (1964). Use of Imperfect Tests for Probability Sampling at the U.S. Bureau of Census. *Bulletin of International Statistical Institute*, 40, 497-516.

Hansen M.H., Hurwitz W.N. and Madow W. G.. (1953). Sample Survey Methods and Theory. *John Willey and Sons, New York, Vol I and II, pp516 ff.*

Hartley, H.O. (1959). Analytical Studies of Survey Data. *Institute of Statistca. Rome* in Honour of Corrode Gini.

Hartley, H.O. (1962). Multiple Frame Surveys. *In proc. Social Statistics Section Amer. Statist. Assoc., Annual Meeting, Minneapolis, Minnesota, 203-206.*

Hartley, H.O. (1974). Multiple Frame Methodology and Selected Applications. Sankhya, C-36, pp- 99-118.

Mahalanobis P. C. (1944). One Large Scale Sample Surveys. *Phil. Trans. R. Soc. – 23, 1(B). pp -329-451*.

Seal K.C. (1962). Use of Out Dated Frames in Large Scale Sample Surveys, *Calcutta Statistical Association Bulletin*, 11, 68-84.

Sen A. R. (1970). On The Bias in Estimation Due to Imperfect Frame in the Canadian Waterfowl Surveys. *The Journal of Wildlife Management. Vol. 34, No. 4 (Oct. 1970), pp. 703-706.*

Singh, N. K. (2020). Frame error in sample survey. *International Research journal of Agricultural Economics and Statistics (IRJAES). Vol.11 (2), Sep., 2020: 240-244.*

Singh, N. K. (2020). Sampling with Imperfect Frame in Large Population. *International Research journal of Agriculture science and research (IJASR)*. Vol.*10, Issue 6, Dec. 2020. 105-112.*

Singh, N. K. (2021). Domain Studies With Imperfect Frame in Large Population. *International journal of Agricultural Sciences. Vol. 17, Issue 2, June 2021, 522-527.*

Singh, N. K., *Rajesh Kumar &* V. K. Sehgal (2001). Use of Incomplete Frame on Large Scale Sample Survey. *Gujarat Statistical Review, Vol. 28 (2001) Nos. 1-2.*

Singh, N.K., Sehgal, V.K. & Kumar, Rajesh (1997). Use of Incomplete Frame for Domain Studies. *Journal of Indian Statistical Association*, 35, 71-81.

Singh, R. (1983), *"On the use of incomplete frame in sample survey"*; Biometrical J, 25, No. 6, 545-549.

Singh, R. (1989). Method of Estimation for Sampling from Incomplete Frames. *Australian journal of statistics, 31 (2), 1989. 269-276.*

Sukhatme, P.V., Sukhatme, B.V. (1984), *Sampling Theory of Surveys with applications*, Iowa State University Press.

Szameitat, K. and Schaffer, K.A. (1963). Imperfect Frames in Statistics and Consequences of Their Use for Sampling, *Bulletin of International statistical Institute*, 40, 517-538.

Turner Anthony G. (2003). Expert Group Meating to Review the Draft Handbook on Designing of Household Sample Surveys. 3-5 December 2003. United Nations Secretariat, Statistical Division. ESA/STAT/AC.93/3.

Yates F. (1948). Sampling Methods for Census and Surveys. *Charles griffin and Co., London* , First Edition.

Yates F. (1960). Sampling Methods for Census and Surveys. *Charles griffin and Co., London* , Third Edition.