CHAPTER - 7

7.1 Sampling with imperfect frame arising due to rare sampling units missing from the target population of very large size:

This chapter deals with the imperfection in the frame arising due rare missing, out-dated or out-of-scope units when the size of the sampled population is very large.

In some situations the size of the sampled population is very large so that it may tend to infinity. Let us consider a situation when size of the population is very large and there are some rare units in the sampled population of very large size, that these are out-dated, out-of-scope or are missing from the target population at the time of actual survey. Such rare missing units not belonging to the desired frame of target population will lead to imperfection in the frame available from sampled population. These rare units of the sampled population of large size missing from the frame desired for the target population under study, will not correspond to the statistical results and observations regarding characteristics of the target population. Quite-often, it happens that in a sampled population of very large size, some rare units missing from the frame will lead to the imperfection in the frame. The deviation between sampled population for which frame is available with large size and the target population for which our results are desired, will contribute to the bias of sampling results. Therefore imperfection in the frame of large size available for the purpose of sample selection will culminate errors in the sampling procedure as unwanted rare units may also be assigned some probability of selection. If it were possible to identify such rare out-of-scope, out-dated or missing units, they could have been assigned zero probability of selection. But in practice it is not possible as frame available at some time regarding sampled population of large size does not have prior information about such rare missing units at the time of sample selection. Therefore frame of large size, available and designed for sample selection becomes imperfect, out-dated and
incomplete because some of the units of sampling frame are missing, out-dated or out-of-scope at the time of actual statistical enquiry due to dynamic nature of population.

Situations arise when frame of large size regarding sampled population, does contain, during passage of time, some rare units which do not belong to the target population at the time of actual survey due to dynamic nature of the sampling frame. Such non-existent units in the target population, if selected in the sample, will not subscribe in accordance with random selection for target population units. But sampler has no option but to select samples from imperfect frame for sampled population of large size and some rare units also get selected, which, infact, do not belong to the target population and which was discovered because such rare non-existent units were selected and investigated by the enumerator.

This may happen because frame of large size prepared or available for some time will, during passage of time, become out-dated by the time of actual survey.

For example in a frame i.e. list of Agricultural holdings of 2 hectares to 5 hectares, available from source or office records regarding year Dec., 1996 in India, the holding sizes may change in rare numbers by the time say March 1997, the actual survey is started. Therefore frame of land holdings becomes imperfect during passage of time. The list of irrigated area, of course large in number, may contain some rare area which in fact is unirrigated and is discovered because that area was selected in the sample and investigated. The list of large size does not have prior information and knowledge of rare units missing from the frame. It is after sampling units are selected at the time of actual survey, enumeration and investigation that units are found to be out-of-scope, out-dated or missing from the target population.
In most of Demographic Studies, frame available for some time from some official sources or other sources are of large size. But some of the demographic units change in time rapidly in their population structure. Thus frame for demographic studies becomes out-dated for time to which data and information relates and the time for which conclusions for statistical results apply, and consequently imperfection in the frame arises.

For example in a study of G.R.R and N.R.R at a time, the frame regarding list of women with reproductive period (15-49 yrs.) change rapidly as some rare number of the women very near to 49 years may be out of reproductive period by the time sample and information is sought for. In a study of fertility rate, some rare women may migrate to other place, hence making the frame imperfect. Thus Demographic units change rapidly in rare number so that they are missing from the frame of large size.

Estimates based on the basis of sample selected from imperfect frame of large size in which rare units are out-dated, out-of-scope or missing units from target population, would not give unbiased results. However it is difficult to get up-to-date and complete frame in large population as rare units of sampled population, most often, cease to exist in the target population by the time actual survey is conducted.

It is assumed that these rare missing units, non-existent in the target population may be of considerable importance and weight for their observations of characteristics under study because of their high measure of size and weight.

Our objective in this chapter is to derive unbiased estimate of target population parameter with imperfect frame arising due to some rare units being out-of-scope, out-dated or missing form the frame so that these are non-existent in the target population of very large size.
7.2  Method of Estimation:

Consider a large sampled population of size \( N \) so that \( N \) tends to infinity. Available frame consists of identifiable, mutually exclusive and non-overlapping units from which sample is to be selected. Therefore size of the frame i.e. \( N \) is so large that it tends to infinity. In a very large variety of investigations, the populations size is so much large than the sample selected from it that it may be regarded as infinite. But due to dynamic nature of population, some rare units from the frame of size \( N \) (\( N \) being large), have undergone some change in the sense that some rare units are non-existent in the target population at the time of actual investigation and survey. Let these rare units missing from the frame are of size \( N_2 \) so that \( N_1 \) units exist in the target population. Sizes of \( N_1 \) and \( N_2 \) are unknown and \( N = N_1 + N_2 \). These \( N_2 \) units are rare in number but due to high measure and weight of each of the rare missing units, they contribute significantly to the population parameter for the character under study. These \( N_2 \) units are not identifiable in the frame and hence cannot be deleted from the frame though they are out-of-scope, out-dated units discovered at the time of investigation and enumeration because such rare units were selected in the sample. Thus actual frame consists of \( N_1 \) units corresponding to the target population which is unknown. Therefore observations of \( N_2 \) rare missing units from the sampled population (of large size \( N \)) will not correspond to the characteristics of the target population. This phenomenon, often happens when population is subject to continuous change for there is long interval between the date to which sampled population relates and date or time for which information and observations are desired for target population. These rare missing units are discovered after these are selected in the sample and enumerated.

The sample selected from such imperfect frame of large size consisting of rare missing units, would not be representative of the target population because the sample may also constitute the rare missing units and randomization principle does not follow for existing
units in the target population. However, if the sampler had prior knowledge of these rare missing units before sample was selected, these rare missing units could have been assigned zero probability of selection and deleted. But we do not know, in advance, about these rare missing units and are discovered because they are selected.

Let \( p \) denote the proportion of rare sampling units which are missing, out-dated and out-of-scope units from the sampling frame of large size. Hence \( p = \frac{N_2}{N} \). Thus \( Np \) units are rare missing units in the sampled population of large size \( N \). Also, \( N_1 \) units actually belong to the target population so that \( N_1 = Nq \), \( q = 1-p \) and \( Np + Nq = N \).

The other notations having the same meaning as elucidated in chapter (VI), the target population total can be written as

\[
Y_{\bar{N}_1} = N_{\bar{N}_1} \bar{Y}_{\bar{N}_1} = Nq \bar{Y}_1
\]

or

\[
\text{Est } Y_{\bar{N}_1} = \text{Est } (Nq \bar{Y}_1)
\]

\[
= \hat{Y}_{\bar{N}_1} \ (\text{say})
\]

or

\[
\hat{Y}_{\bar{N}_1} = Nq \bar{Y}_1 \text{ as Est } \bar{Y}_1 = \bar{y}_1 \text{ and Est } q = \hat{q}.
\]

Now in order to obtain the estimator of target population total for \( y \), we have to estimate \( q \).
7.3 Sampling Scheme :-

In the population of large size N, the number of missing or out-dated units are rare in number. Therefore proportion p is very small. In this case also, method of sample selection as given in chapter (II) would not be appropriate. The method of Inverse-sampling, however, can be used with advantage because even a large sample with S.R.S.W.O.R. may not provide representation to rare missing units from the large population. Some predetermined number of rare units missing from the frame should be included in the sample so as to give representation to such out-dated and out-of-scope units for obtaining unbiased results.

Following the same procedure of I-SRSWOR as given in chapter (VI), the sample size n is not fixed in advance. Instead sampling is continued until a predetermined number of rare units missing from the frame of large size, have been drawn. Let, our statistical enquiry requires that the sample should include $n_2$ units from the rare missing units of frame with large size N which may tend to infinity.

Following the same procedure of I-SRSWOR we continue selecting units one by one, with equal probability and without replacement, from the sampled population $U = \{U_1, U_2, U_3, \ldots, U_n\}$; until there are exactly $n_2$ units (given) discovered at the enumeration stage represented from the rare missing units. Thus n is a random variable. The sample size n will contain $n_2$ rare units which do not belong to the target population and $n-n_2 = n_1$ (say) units discovered as belonging to the target population, after each unit of random size n is investigated, at the time of actual enumeration.

Therefore no observations can be obtained for $n_2$ rare out-dated, out-of-scope or missing unit as discovered. Even if observations are available, they will not contribute and correspond to the target population characteristics because investigator has discovered and identified them as not belonging to the target population and hence can not be observed.
Let $P(n)$ be the probability that exactly $n$ units are required before $n_2$ rare missing, out-of-scope, or out-dated units are observed in the sample. The first $n-1$ units must include $n_2-1$ rare missing unit from the frame and the $n^{th}$ unit is the rare missing unit in the frame. Therefore, for large population of size $N$, we have

$$P(n) = P\left\{ \frac{n-1}{n_2-1} \text{ units include rare missing}, \frac{n}{n_2-1} \text{ the rare missing unit} \right\}$$

or

$$P(n) = \left( \frac{n-1}{n_2-1} \right) p^{n_2-1} q^{n-n_2}.$$ 

or

$$P(n) = \left( \frac{n-1}{n_2-1} \right) p^{n_2} q^{n-n_2}.$$ 

$$n = n_2, n_2 + 1, n_2 + 2, \ldots \ldots .$$

This is coefficient of $t^n$ in $(pt/1-qt)^{n_2}$

This is negative Binomial distribution.

Putting $r=n-n_2$ we can have

$$P(n) = \left( \frac{n_2+r-1}{r} \right) p^{n_2} q^r$$

for

$$\binom{n_2+r-1}{n_2-1} = \binom{n_2+r-1}{r}$$

$$r = 0, 1, 2, 3, \ldots \ldots$$

so that $\sum_{r=0}^{\infty} P(n) = 1$
7.4 Estimation of proportion $p$ and its variance:

It can be shown that an unbiased estimate of $p$ can be given as

$$ p = \frac{n_2 - 1}{n - 1} \quad (7.4.1) $$

Because

$$ E(\hat{p}) = E \left( \frac{n_2 - 1}{n - 1} \right) $$

$$ = \sum_{n \geq n_2} \left( \frac{n_2 - 1}{n - 1} \right) P(n) $$

$$ = \sum_{n \geq n_2} \left( \frac{n_2 - 1}{n - 1} \right) \left( \frac{n - 1}{n_2 - 1} \right) n_2^{n-n_2} p^n q^{n-n_2} $$

$$ = \sum_{n \geq n_2} \left( \frac{n - 2}{n_2 - 2} \right) n_2^{n-n_2} p^n q^{n-n_2} $$

$$ = \sum_{n \geq n_2} \left( \frac{(n-1)-1}{(n_2-1)-1} \right) n_2^{n-1} n^{n-n_2} p^n q^{n-n_2} $$

$$ = p \quad \text{as terms of the summation equals to 1.} $$

Also, by putting $n - n_2 = r$ we can have

$$ E(\hat{p}) = \sum_{n \geq n_2} \left( \frac{n-2}{n_2-2} \right) n_2^{n-n_2} p^n q^n $$

$$ = p^2 \sum_{r \geq 0} \left( \frac{n_2 + r - 2}{n - r - 1} \right) q^r $$

$$ = p^2 \sum_{r \geq 0} \left( \frac{n_2 + r - 2}{r} \right) q^r $$
\[ p^2 (1-q)^{1-n_2} = p \]

Now we shall determine the estimate of \( V(\hat{p}) \).

We know that
\[ V(\hat{p}) = E(\hat{p}^2) - \{E(\hat{p})\}^2 \]
\[ = E(\hat{p}^2) - p^2 \quad (7.4.2) \]

Therefore an unbiased estimate of \( V(\hat{p}) \) is given by
\[ \text{Est } V(\hat{p}) = \hat{p}^2 - \text{Est } p^2 = \hat{V}(\hat{p}), \text{ (say)} \quad (7.4.3) \]

Now to determine \( \text{Est } p^2 \) we have

\[ E\left(\frac{(n_2 - 1)(n_2 - 2)}{(n-1)(n-2)}\right) = \sum_{n \geq n_2} \frac{(n_2 - 1)(n_2 - 2)}{(n-1)(n-2)} \binom{n-1}{n_2-1} p^{n_2} q^{n-n_2} \]

\[ = p^2 \sum_{n \geq n_2} \binom{n-3}{n_2-3} p^{n_2-2} q^{n-n_2} \]

\[ = p^2 \sum_{n \geq n_2} \binom{n_2-2-1}{n_2-2} p^{n_2-2} q^{n-n_2} \]

\[ = p^2 \quad (7.4.4) \]

Therefore
\[ \text{Est } p^2 = \frac{(n_2 - 1)(n_2 - 2)}{(n-1)(n-2)} \quad (7.4.5) \]

Therefore from (7.4.3) and (7.4.5) we have
\[
\text{Est } V(p) = \hat{p}^2 - \left( \frac{n_2 - 1}{n-1} \right) \left( \frac{n_2 - 2}{n-2} \right) \\
= \hat{p} \left( \hat{p} - \frac{n_2 - 2}{n-2} \right) \quad \text{as } \hat{p} = \frac{n_2 - 1}{n-1} \\
= \frac{(n_2 - 1)}{n-1} \left\{ \frac{n_2 - 1}{n-1} - \frac{n_2 - 2}{n-2} \right\} \\
= \frac{n_2 - 1}{n-1} \left\{ \frac{n-n_2}{(n-1)(n-2)} \right\} \\
\]

or
\[
\hat{V}(p) = \frac{(n_2 - 1)(n-n_2)}{(n-1)^2(n-2)} \tag{7.4.6}
\]

Also
\[
\text{Est } V(p) = \frac{(n_2 - 1)}{(n-1)(n-2)} \left\{ \frac{n-n_2}{n-1} \right\} \\
= \left( \frac{n_2 - 1}{n-1} \right) \left( \frac{1}{n-2} \right) \left( 1 - \frac{n_2 - 1}{n-1} \right) \\
= \frac{\hat{p}(1-\hat{p})}{n-2} \tag{7.4.7}
\]
7.5 Unbiased Estimate of target population total :-

Now unbiased estimate of target population total can be given as, \( \hat{Y}_{N_1} \) (say), so that

\[
\text{Est } Y_{N_1} = \text{Est } (Nq \bar{Y}_1) \\
= N \text{ Est } (q \bar{Y}_1) \\
= N \text{ Est } q \text{ Est } \bar{Y}_1 \\
= N q \bar{Y}_1 \\
= N (1-p) \bar{Y}_1 \quad \text{as } p + q = 1 \\
= N \left( 1 - \frac{n^2 - 1}{n-1} \right) \bar{Y}_1 \\
= N \left( \frac{n - n^2}{n - 1} \right) \bar{Y}_1 \\
(7.5.1)
\]

\( \hat{Y}_{N_1} \) is an unbiased estimate of \( Y_{N_1} \) because

\[
E(\hat{Y}_{N_1}) = E(Nq \bar{Y}_1) \\
= N E(q) E(\bar{Y}_1) \quad \text{as } E(\bar{Y}_1) = \bar{Y}_1 \\
= N (1-p) \bar{Y}_1 \\
= N (1-p) \bar{Y}_1 \\
= N q \bar{Y}_1 \quad \text{i.e. target population total} \\
= N \bar{Y}_{N_1} \quad \text{as } Nq = N_1
\]

133
7.6 Estimation of the Variance of $\hat{Y}_{N1}$

Now we shall determine estimate of the variance of the estimate of target population total. The notations have the same meaning as given in chapter VI.

Let us have variance of $\hat{Y}_{N1}$ as follows

$$V(\hat{y}_{N1}) = V(Nq \bar{y}_{1})$$

$$= N^2 V \{(1-p) \bar{y}_{1}\}$$

$$= N^2 V (\bar{y}_{1} - p \bar{y}_{1})$$

$$= N^2 \left[ V(\bar{y}_{1}) + V(p \bar{y}_{1}) \right]$$

$$= N^2 \left[ V(\bar{y}_{1}) + E \left(p^2 \bar{y}_{1}^2 + V(p) \right) \right]$$

But we have $E(p^2) = V(p) + p^2$

and $\bar{y}_{1}^2 = E(\bar{y}_{1}^2) - V(\bar{y}_{1})$

Therefore

$$V(\hat{y}_{N1}) = N^2 \left[ V(\bar{y}_{1}) + V(p \bar{y}_{1}) \right]$$

$$= N^2 \left[ V(\bar{y}_{1}) (1+p^2) + V(p) E(\bar{y}_{1}^2) \right]$$

(7.6.1)

Therefore estimate of $V(\hat{y}_{N1})$ can be given as

$$Est \ V(\hat{y}_{N1}) = N^2 \left[ Est \left( V(\bar{y}_{1})(1+p^2) \right) + Est \left( V(p) E(\bar{y}_{1}^2) \right) \right]$$

$$= N^2 \left[ V(\bar{y}_{1}) Est (1+p^2) + V(p)\bar{y}_{1}^2 \right]$$

(7.6.2)

134
But \( \text{Est } (1+p^2) = 1 + \text{Est } p^2 \)

Where \( \text{Est } p^2 = \frac{(n_2 - 1)(n_2 - 2)}{(n-1)(n-2)} \) and \( \hat{V}(\hat{y}_1) = \frac{N_1 - n_1}{N_1 n_1} \frac{s^2}{n_1} \)

Therefore we can have

\[
\hat{V}(\hat{Y}_{N_1}) = N^2 \left[ \left\{ \frac{N_1 - n_1}{N_1 n_1} s^2 \right\} \left\{ 1 + \frac{(n_2 - 1)(n_2 - 2)}{(n-1)(n-2)} \right\} \right] \\
+ \hat{y}_1^2 \left\{ \frac{(n_2 - 1)(n-n_2)}{(n-1)^2 (n-2)} \right\} 
\]  

(7.6.3)

or

\[
\hat{V}(\hat{Y}_{N_1}) = N^2 \left[ \left\{ \frac{N_1 - n_1}{N_1 n_1} s^2 \right\} \left\{ 1 + \frac{p^2 - \hat{p}(1-p)}{n-2} \right\} \right] \\
+ \hat{y}_1^2 \left\{ \frac{\hat{p} (1-\hat{p})}{n-2} \right\} 
\]  

(7.6.4)

### 7.7 Estimation of \( 1/N_1 \):

As we know that value of \( N_1 \) is unknown because sampler does not know, in advance that how many sampling units belong in the target population. Therefore \( N_1 \) has to be inserted with its estimate determined on the basis of sample.

Again \( \frac{N_1 - n_1}{N_1 n_1} s^2 = \left( \frac{1}{n_1} - \frac{1}{N_1} \right) s^2 \)

Therefore \( \text{Est } \left( \frac{N_1 - n_1}{N_1 n_1} \right) = \frac{1}{n_1} - \text{Est } \left( \frac{1}{N_1} \right) \).

Again assuming \( \hat{N}_1 = N_1 + \epsilon \), \( E(\epsilon) = 0 \) we have

\( E\left( \frac{1}{\hat{N}_1} \right) = E\left( \frac{1}{N_1} \right) \) and hence \( \text{Est } \left( \frac{1}{N_1} \right) = \frac{1}{N_1} \)
Where \( \hat{N}_1 = \text{Est} (Nq) = \hat{N}_q = N(1-\hat{p}) \).

Putting these values in (7.6.3) we have

\[
\hat{V}(\hat{Y}_{N_1}) = N \left\{ \frac{N(1-\hat{p})-n_1}{(1-\hat{p})n_1} \right\} \hat{s}_1^2 \left\{ 1 + \frac{(n_2-1)(n_2-2)}{(n-1)(n-2)} \right\} 
+ N^2 \hat{y}_1^2 \left\{ \frac{(n_2-1)(n-n_1)}{(n-1)^2(n-2)} \right\} 
\]  

(7.7.1)

Now as \( 1-\hat{p} = 1 - \frac{n_2-1}{n_2} = \frac{n-n_2}{n-1} = \frac{n_1}{n-1} \) for \( n_1 + n_2 = n \)

After simplification we have

\[
\hat{V}(\hat{Y}_{N_1}) = \frac{N(N-n+1)}{n_1} \hat{s}_1^2 \left\{ 1 + \frac{(n_2-1)(n_2-2)}{(n-1)(n-2)} \right\} 
+ N^2 \hat{y}_1^2 \left\{ \frac{(n_2-1)(n-n_1)}{(n-1)^2(n-2)} \right\} 
\]  

(7.7.2)

Thus we see that \( \hat{V}(\hat{Y}_{N_1}) \) is function of \( N, n_1, n_2, n, \hat{s}_1^2 \) and \( \hat{y}_1^2 \). Thus values can be easily obtained on the basis of the sample observations. Therefore an unbiased estimate of variance of estimate of the target population can be obtained with imperfect frame arising due to rare missing units in the large population. It is interesting to note that as \( n_1 \) increases, the \( \hat{V}(\hat{Y}_{N_1}) \) also increases because of \( n-n_2=n_1 \) that occurs in second term. First term has multiplier \( N(N-n+1)/n_1 \) and second term includes \( N^2 \hat{y}_1^2 \).

Hence \( \hat{V}(\hat{Y}_{N_1}) \) increases as \( n_1 \) increases. It is in the case of size of sampled population being very large. However in case of finite sampled population as explained in chapter VI
\( \hat{V}(\hat{Y}_{N_1}) \) decreases as \( n_1 \) increases. Here also if \( \hat{p} = 1 \) so that \( n = n_2 \) and \( n_1 = 0 \) then \( \hat{V}(\hat{Y}_{N_1}) \) cannot be estimated. Again if \( n_1 = 1 \) then \( s_1^2 = 0 \) and we get

\[
\hat{V}(\hat{Y}_{N_1}) = N^2 y_1^2 \left( \frac{n_2 - 1}{(n-1)^2} \right) \]

\[
= \frac{N^2 y_1^2}{(n-1)^2} \text{ because } n_2 - 1 = n - 2
\]

where \( y_1 \) is single observation and \( \bar{y}_1 = y_1 \)

Again if \( \hat{p} = 0 \), the situation when there is no imperfection in the frame and every unit of the sampled population correspond with each identifiable unit of the target population, we can obtain

\[
\hat{V}(\hat{Y}_{N_1}) = \frac{N (N-n+1)}{n} s_1^2 \quad \text{for } n_1 = n \quad \text{if } \hat{p} = 0
\]

or we can have \( N-n+1 \approx N-n \) so that

\[
\hat{V}(\hat{Y}_{N_1}) = \frac{N(N-n_1)}{n_1} s_1^2 = \frac{N(N-n)}{n} s^2
\]

Since for \( n = n_1 \), \( s_1^2 = s^2 \).

This result attributes to the ideal condition which can easily be obtained for no imperfection in the frame of large size with S.R.S.W.O.R.