UNIT 14 CLUSTER SAMPLING AND MULTISTAGE SAMPLING

Structure **Page No.**

14.1 INTRODUCTION

In all the sample selection procedures discussed so far in this block, the entire investigation was based on the assumption that a usable list of units (i.e., a *frame*) is available from which one selects a sample. Unfortunately, not always such a list of units is available especially when we are concerned with *countrywide investigations*. Even existence of such a list of units under investigation (in some cases) would not give us enough scope to base our enquiry on a simple random sample because of high budgeting costs.

For example, if the sampling units are individuals, a random sample is likely to be scattered evenly over the region under survey making it difficult to conduct survey with low cost. Alternatively, for example, if *chunks* of households or villages are selected as sampling units then these units can be *clustered* to make the survey more cost effective, particularly with respect to travel cost.

As an example from agriculture, if the ultimate units of observation are fields, a sampling unit may as well be a larger collection (i.e., **cluster**) of fields such as area units like villages or segments.

Also, as the efficiency of estimators depends on the fact whether the sampling units are clusters of individuals or the individuals themselves, the *choice of sampling unit becomes an important consideration for a proper sampling plan*. And, the problem of frames can be handled more effectively by forming *clusters of basic sampling units*.

The procedure of selecting **clusters** and then observing all the elements in the selected clusters is known as **cluster sampling**. A natural extension of idea of cluster sampling is *sub-sampling* in which the clusters selected at a later stage are further sub-sampled. The procedure of sub-sampling can be extended to **multi-stage sampling**.

In Sec.14.2, we shall talk about certain preliminary aspects of *cluster sampling*, discuss relations used in the *estimation of population mean*, and describe briefly the *efficiency of cluster sampling*. In Sec.14.3, after having discussed certain preliminary aspects of *multi-stage sampling*, we shall discuss another set of relations used in the estimation of population mean in the context of *two-stage sampling*.

Objectives

After reading this unit, you should be able to

- discuss a situation for using cluster/multistage sampling;
- estimate the population mean in case of equal and unequal size of clusters;
- estimate the relative efficiency of cluster sampling;

- differentiate between cluster sampling and two-stage sampling;
- estimate the population mean in case of two-stage sampling.

14.2 CLUSTER SAMPLING

As said in the introduction, when the sampling unit is a cluster, the procedure of sampling is called cluster sampling. *So,* **cluster sampling** *consists of forming suitable clusters of contiguous population units and surveying all the units in a* **sample of clusters** *selected according to some appropriate sample selection method.*

For instance, consider a big village as a cluster of farmers. Then, for selecting farmers from the area, certain smaller villages may be selected and information from farmers of these villages is obtained. Here, it is important to mention that a *list of farmers* in a region may not be available but a *list of villages* is always available. This example is typical in *area sampling*.

Other then the *area sampling*, there are situations when cluster sampling is of great help. For example, if one wishes to interview passengers departing from an airport, then a *cluster might be the plane load.* On the other hand, if one is searching through files of land holdings for tax information, then *pages in a ladger* would be the *clusters.*

There are situations when conducting a survey with clusters of sampling units, instead of taking a simple random sample from a population, is cost effective.

For instance, if a sample is selected from the population of all *sixth-grade* students in a particular state, then each school in the state is taken as a cluster of the basic sampling units and we choose a simple random sample of a few schools and interview all the *sixth-graders* in those schools according to pre-set survey objectives. However, a simple random sample of 400 students usually, as we may agree to, better represents the entire population – and therefore provides better information about the population – than a group of 100 students studied in each of the four specified schools.

Thus, in general, the choice for a sampling procedure to be adopted should be guided by cost considerations and by the degree of precision desired in estimating the population parameters.

Also, it is important to realise that the cluster sampling in above situation has actualy helped in avoiding the necessity of constructing a *frame* for the entire population, which is certainly an exhausting and expensive job in itself. In addition to that, cluster sampling is remarkably expedient because the units in a cluster are adjacent and therefore easy to approach.

Now, let us talk about some more aspects of cluster sampling in detail. We start by talking about some introductory aspects of cluster sampling.

14.2.1 Preliminaries

Let us consider a case of cluster sampling in which a number of people in a city are to be interviewed. For selecting a sample, the telehpone directories are used and it is decided to interview people through telephone. Now, since all the residents can be numbered, a random sampling technique could have been used to choose sample houses.

Also, we could form strata of houses for *high, middle,* and *low* income groups. Now, if we choose houses throughout the city in random manner, then cost of visiting widely scattered dwellings will certainly be prohibitive.

An alternative way of sample selection is to *group blocks* or *areas* into clusters of approximately equal population. Then, a number of these clusters can be chosen at

random. Within each cluster, all households may be interviewed. On comparing this (cluster) sampling procedure with that of making random choice of households throughout the city, it is clear that the cost per element (a household) is certainly going to be lower because of lower listing cost (as it is necessary only to list the houses on the blocks selected) and lower location cost. Also, it is going to be easy for an interviewer to talk to several people on one block rather than to several people scattered throughtout the city.

Note that a necessary condition for the validity of above procedure is that every unit of the population under study must correspond to one and only one unit of the cluster so that the total number of sampling units in the list (frame) will cover all the units of the population under study with no omission or duplication. When this condition is not satisfied, estimators become **biased**.

In the following table, we fix some notations for our convenience and future use in this unit. We shall use these notations ferquently while calculating estimators of population parameters.

Table 1. Notations used in this section and their meanings.

N	number of clusters in a population.
n	number of clusters in the sample.
$M_i, 1 \leq i \leq N$	number of units in the i-th cluster of a population.
$M_0 = \sum_{i=1}^{N} M_i$	total number of units in a population.
$\overline{M} = \frac{M_0}{N}$	average number of units per cluster in a population.
Y_{ij}	value of the character under study for the j-th unit in the i-th cluster, $j = 1, 2, \ldots, M_i;\ i = 1, 2, \ldots, N$.
$Y_i = \sum_{j=1}^{M_i} Y_{ij}$	i-th cluster total.
$Y = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} Y_i$	total of Y-values for all the M_0 units in a population.
$\overline{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij} = \frac{Y_i}{M_i}$	the mean per unit of the i-th cluster.
$y_k = \sum_{k=1}^{M_k} Y_{kj} (= Y_k)$	k-th sample cluster total ($1 \leq k \leq n$).
$\overline{Y}_c = \frac{Y}{N} = \frac{1}{N} \sum_{i=1}^{N} Y_i$	population mean per cluster.
$\overline{Y} = \frac{\sum_{i=1}^{N} M_i \overline{Y}_i}{M_0} = \frac{Y}{M_0}$	population mean per unit.

Let us workout a problem to get familiar to the notations defined in Table 1.

Problem 1. Suppose from a total of 20 bearing trees of guaua in a village, 5 clusters of size 4 trees each were selected and (hypothetical) yield (in kgs) is as given in the following table.

cluster	*1st* tree	*2nd* tree	*3rd* tree	*4th* tree
1	5	4	1	15
2	11	1	4	7
3	36	10	19	11
4	7	15	12	10
5	2	22	8	6

Claculate the quantities $Y_i, \overline{Y}_i$ $(1 \leq i \leq 5), \overline{Y}_c$ and $\overline{Y}$.

Solution. Here, $n = 5, M_i = 4$, for all i, and $N = 20$. Then, using values of Y_{ij} from above table, we get

$$Y_1 = \sum_{j=1}^{M_i} Y_{1j} = \sum_{j=1}^{4} Y_{1j} = 5 + 4 + 1 + 15 = 25, \text{ and}$$

$$Y_2 = \sum_{j=1}^{4} Y_{2j} = 11 + 1 + 4 + 7 = 23.$$

Similarly, you can find that $Y_3 = 76, Y_4 = 4438$ and $Y_5 = 38$. Thus,

$$\overline{Y}_1 = \frac{Y_1}{M_1} = \frac{25}{4}, \overline{Y}_2 = \frac{Y_2}{M_2} = \frac{23}{4}, \overline{Y}_3 = \frac{76}{4}, \overline{Y}_4 = \frac{44}{4} \text{ and } \overline{Y}_5 = \frac{38}{4}.$$

Finally,

$$\overline{Y}_c = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{1}{5}(25 + 23 + 76 + 44 + 38) = 41.2, \text{ and}$$

$$\overline{Y} = \frac{Y}{M_0} = \frac{\sum_{i=1}^{5} Y_i}{M_0} = \frac{25 + 23 + 76 + 44 + 38}{20} = 10.3.$$

Now you try to solve the following exercise.

E1) From your daily life experiences, give five examples where cluster sampling can be used.

As you may have observed, the clusters are usually formed by grouping neighbouring units or units which can be conveniently surveyed together. The construction of clusters, however, differs from the optimal construction of strata that you read about in Unit 13.

For example, household is a cluster of individuals, village is a cluster of farmers and a class, which is a group of students, is a cluster. Similarly, an orchard can be considered as cluster of trees, etc.

Generally, while stratification may reduce sampling error, clustering tends to decrease costs and increase sampling error for the same size of sample. This is mainly because people who live close together are more likely to be similar than those living in different localities.

As you read in previous unit, while stratifying a population, strata are as homogeneous as possible within themselves and differ as much as possible from each another with respect to the study variable. And, units within a stratum need not be geographically contiguous. On the other hand, if an estimate based on cluster sampling has to be more efficient than the estimates made with simple random sampling procedure, clusters should be internally as *heterogeneous* as possible.

For a given total number of units in the sample, the cluster sampling is usually less efficient than sampling of individual units as the latter is likely to provide a better cross section of the population units than the former. This is essentially due to tendency of units in a cluster to be similar. Another fact that we would like to share with you is that the efficiency of cluster sampling is likely to decrease with increase in cluster size.

All said and done, cluster sampling is operationally convenient and economical than

Some of the major Government agencies, Universities research studies, and marketing research use a combination of clustering and stratification techniques to control cost and error and to provide adequate size groups for intensive studies.

sampling of individual units. In many practical situations, the loss in efficiency in terms of sampling variance is likely to be balanced by the reduction in cost particularly the travel cost between units.

Hence, because of its operational convenience and possible reduction in cost, the survey tasks in many situations are facilitated by using nonoverlapping and collectively exhaustive cluster of units. Now, in the next part of the section, we shall discuss some relations used in the estimation of population parameters.

14.2.2 Estimation of Population Mean

Simple random sampling, systematic sampling, and stratified sampling are various types of sampling procedures that can be applied in the cluster sampling by treating the clusters as sampling units. In this unit, however, we shall restrict to situations where clusters are selected using *without replacement simple random sampling* procedure.

The theory of cluster sampling in its own right is rather complex, where the complexity depends on whether one takes *equal* or *unequal-sized* clusters. In general, a formulae for calculating the *standard error* of cluster estimates has two terms, where the first relates to the variability between cluster means (or proportions) and the second to the variability within cluster.

In this unit, we start with the case of unequal clusters and then deduce from this the results about clusters of equal sizes as a special case.

Case-I: Unequal clusters. Usually, in practice, clusters are of unequal sizes. For instance, households as a group of persons and villages as a group of households can be taken as clusters for the purpose of sampling.

We assume (i) the population consists of N clusters, where the ith cluster has M_i elements, $i = 1, 2, \ldots, N$, and (ii) n clusters are selected from N clusters by *without replacement simple random sampling* procedure.

Throughout, the *suffix* **c** refers to **cluster sampling**.

Then, an unbiased estimator of population mean $\overline{Y}$, with M_0 known, is given by the relations

$$\begin{aligned}\hat{\overline{Y}}_c &= \frac{N}{nM_0}\sum_{k=1}^{n} M_k \overline{Y}_k \\ &= \frac{1}{\overline{M}n}\sum_{k=1}^{n} Y_k, \text{ where } \overline{M} = \frac{M_0}{N}.\end{aligned}$$

And, the variance of the estimator $\hat{\overline{Y}}_c$ is given by the relation

$$V(\hat{\overline{Y}}_c) = \left(\frac{N-n}{N\,n\,\overline{M}^2}\right)\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y}_c)^2.$$

Also, an unbaised estimator of variance is given by relations

$$\begin{aligned}\hat{V}(\hat{\overline{Y}}_c) &= \left(\frac{N-n}{N\,n\,\overline{M}^2}\right)\frac{1}{n-1}\sum_{k=1}^{n}\left(Y_k - \overline{M}\,\hat{\overline{Y}}_c\right)^2 \\ &= \left(\frac{N-n}{N\,n\,\overline{M}^2}\right)\frac{1}{n-1}\left[\left(\sum_{k=1}^{n} Y_k^2\right) - n(\overline{M}\,\hat{\overline{Y}}_c)^2\right]\end{aligned}$$

We try to understand a use of these relations with the help of the following problem.

Problem 2. For studying the cultivation practices and yield of apple, a pilot sample survey is conducted in a district of Kashmir. The yield (in kgs) of 3 clusters of trees, selected by without replacement simple random sampling, from 15 are as given in the following table.

cluster	size	yield
1	12	5.53,26.11,11.08,12.66,0.87,6.40,54.31,37.94, 7.13,3.53,14.23,1.24
2	10	4.84,10.93,0.65,32.52,3.56,11.68 35.97,47.07,17.69,40.7
3	6	15.79,11.18,27.54,28.11,21.70,1.25

With $\overline{M} = 10$, estimate the average yield per tree as well as the production of apple in the village and their standard errors.

Solution. Here, $N = 15, n = 3$ and $\overline{M} = 10$. Then, $M_0 = N\overline{M} = 150$. So, $M_1 = 12, M_2 = 10$ and $M_3 = 6$. Then, using values from table, we get

$$\hat{\overline{Y}}_c = \frac{1}{30}[Y_1 + Y_2 + Y_3] = \frac{1}{30}\left[\sum_{j=1}^{M_1} Y_{1j} + \sum_{j=1}^{M_2} Y_{2j} + \sum_{j=1}^{M_3} Y_{3j}\right]$$
$$= \frac{1}{30}[181.03 + 205.61 + 105.57] = \frac{492.21}{30} = 16.407.$$

Thus, the average yield of apple per tree is 16.407 kgs. Also, we have

$$Y_1 = \sum_{j=1}^{M_1} Y_{1j} = 181.03,\ Y_2 = \sum_{j=1}^{M_2} Y_{2j} = 205.61, \text{ and } Y_3 = \sum_{j=1}^{M_3} Y_{3j} = 105.57.$$

Using these values of $Y_i (1 \le i \le 3)$, the estimated variance $\hat{V}(\hat{\overline{Y}}_c)$ is given by

$$\hat{V}(\hat{\overline{Y}}_c) = \left(\frac{15-3}{15 \times 3 \times 100}\right)\frac{1}{3-1}\left[\left(\sum_{k=1}^{n} Y_k^2\right) - 3(10 \times 19.709)\right]$$
$$= \frac{1}{750}[(32771.86 + 42275.47 + 11145.03) - 26918.97] = 79.03$$

Thus the standard error of $\hat{\overline{Y}}_c$ is $\sqrt{79.03} = 8.89$.

———— × ————

Case-II: Equal clusters. Let $M_i = M$, for all i. That is, unequal clusters reduce to clusters of equal sizes. Hence, in case of equal clusters, the unbiased estimator of population mean is given by

$$\hat{\overline{Y}}_c = \frac{1}{n}\sum_{k=1}^{n} \overline{Y}_k, \text{ as } M_0 = NM, \text{ in this case.}$$

Similarly, the variance of the estimator $\hat{\overline{Y}}_c$ is now given by

$$V(\hat{\overline{Y}}_c) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2$$
$$\cong \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \{1 + (M-1)\rho\}, \text{ for large N,}$$

where ρ is *intra-cluster correlation coefficient* between elements within clusters, S^2 is population mean square and (with $\overline{M} = M$)

ρ is a measure of *internal homogeneity* of the clusters.

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\overline{Y}_i - \overline{Y})^2.$$

Generally, ρ is positive since clusters are usually found by putting together geographical contiguous *farms, stores, establishments, families*, etc. Thus, for the same number of units in a sample, cluster sampling gives a higher variance than sampling elements directly.

But the real point here is that it is far cheaper to collect information on a *per-unit basis* if sampling is done in clusters. If ρ is negative, both cost and the variance suggest a use of clusters. Furthermore, an unbiased estimator of $V(\hat{\overline{Y}}_c)$ in this situation is given by

$$\hat{V}(\hat{\overline{Y}}_c) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2, \text{ where } s_b^2 = \frac{1}{n-1}\sum_{k=1}^{n} \left(\overline{Y}_k - \hat{\overline{Y}}_c\right)^2$$

Once again, we shall try to understand these relations with help of a practical situation.

Problem 3. A pilot sample survey was conducted to study the management practices and yields of apple in a village of Himachal Pradesh (India). Of the total 300 bearing trees, 10 clusters of size 3 each were selected and their yield records (in kgs) are as given in the following table.

cluster	*1st* tree	*2nd* tree	*3rd* tree
1	6.52	5.73	15.24
2	12.08	1.65	9.28
3	24.15	30.75	17.26
4	16.24	8.20	6.58
5	54.92	34.62	12.16
6	36.24	46.28	28.54
7	42.48	36.34	26.42
8	18.24	16.80	91.46
9	54.27	40.28	25.55
10	1.94	5.16	22.70

Estimate the average yield per tree along with its standard error.

Solution. Here, $M_0 = 300, n = 10$ and $M_i = 3$, for $i = 1, 2, \ldots, 10$. So, proceeding as in Problem 2, you can see that the average yield per tree is $\hat{\bar{Y}}_c = 24.94$. Also, the estimated variance $\hat{V}(\hat{\bar{Y}}_c)$, with $N = \frac{M_0}{M} = 100$, is given by

$$\begin{aligned}
\hat{V}(\hat{\bar{Y}}_c) &= \left(\frac{1}{10} - \frac{1}{100}\right) s_b^2 = \frac{9}{100} \times \frac{1}{9} \times \left(\sum_{k=1}^{10} (\bar{Y}_k - \hat{\bar{Y}}_c)^2\right) \\
&= \frac{1}{100}\left[\left(\frac{Y_1}{M_1} - \hat{\bar{Y}}_c\right)^2 + \left(\frac{Y_2}{M_2} - \hat{\bar{Y}}_c\right)^2 + \ldots + \left(\frac{Y_{10}}{M_{10}} - \hat{\bar{Y}}_c\right)^2\right] \\
&= \frac{1}{9 \times 100}\left[\left(Y_1 - 3\hat{\bar{Y}}_c\right)^2 + \left(Y_2 - 3\hat{\bar{Y}}_c\right)^2 + \ldots + \left(Y_{10} - 3\hat{\bar{Y}}_c\right)^2\right] \\
&= \frac{1}{900}\left[(27.49 - 74.81)^2 + (23.01 - 74.81) + \ldots + (29.8 - 74.81)^2\right] \\
&= 18.399.
\end{aligned}$$

Thus, the standard error is $\sqrt{18.399} = 4.299$.

———— × ————

Now, you try the following exercise.

E2) Change some of the figures in the last three columns of the table given in Problem 3 above, and then calculate the average yield per tree along with its standard error.

Above we obtained expressions for an unbiased estimator of population mean alongwith expressions for its variance and estimator of variance. In the next part of the section, we shall describe the relative efficiency aspect of cluster sampling. For this purpose, we shall assume situations where all clusters are of equal size.

14.2.3 Efficiency of Cluster Sampling

Here, right in the beginning, we want to remark that the estimator $\hat{\bar{Y}}_c$ for equal sized clusters is based on a sample of nM units in the form of n clusters each consisting of M units. Thus, if the same number of units are selected from a population of NM units by *without replacement simple random sampling* procedure, then the sample mean estimator $\hat{\bar{Y}}$ and its variance $V(\hat{\bar{Y}})$ are given by the relations (see Unit 12)

$$\hat{\bar{Y}} = \frac{1}{nM}\sum_{k=1}^{nM} Y_k, \text{ and}$$

$$V(\hat{\bar{Y}}) = \left(\frac{1}{n\,M} - \frac{1}{N\,M}\right) S^2$$

$$= \left(\frac{N-n}{N\,n\,M}\right) \frac{1}{N\,M-1} \left(\left(\sum_{i=1}^{NM} Y_i^2\right) - NM\,\bar{\bar{Y}}^2\right), \text{ respectively.}$$

And, in relation to the sample mean estimator $\hat{\bar{Y}}$, the *relative efficiency* (**RE**, in short) of the estimator $\hat{\bar{Y}}_c$ for equal sized clusters is given by $RE = \dfrac{V(\hat{\bar{Y}})}{V(\hat{\bar{Y}}_c)}$, where $V(\hat{\bar{Y}}_c)$ denotes the variance for equal sized cluster.

Observe that the relative efficiency defined above involves value of study variable for all population units. However, in practice, the investigator has only the sample observations of n clusters of M units each. For this, he needs the estimates of two variances involved in the formulae of relative efficiency (RE).

An unbiased estimator of $V(\hat{\bar{Y}})$ from a cluster sample is given by

$$\hat{V}(\hat{\bar{Y}}) = \frac{N-n}{(NM-1)n} \left[\left(\frac{1}{n\,M} \sum_{k=1}^{n} \sum_{j=1}^{M} y_{kj}^2\right) + \hat{V}(\hat{\bar{Y}}_c) - \hat{\bar{Y}}_c^2\right],$$

while an unbiased estimator of $V(\hat{\bar{Y}}_c)$ for equal size clusters is same as given in the previous part of the section. Then, *estimator of relative efficiency* $\hat{RE}$ of estimator $\hat{\bar{Y}}_c$ (for equal size clusters) with respect to the usual estimator $\hat{\bar{Y}}$ from a cluster sample is given by $\hat{RE} = \dfrac{\hat{V}(\hat{\bar{Y}})}{\hat{V}(\hat{\bar{Y}}_c)}$.

Let us now discuss a practical situation to see how above stated relation are used in practice.

Problem 4. A company has 25 centres located at different places in a State. Each centre has been provided with 4 telephones. In order to estimate the average number of calls per telephone made on a typical day for this company, a sample of 5 centres, using without replacement simple random sampling, were selected. The data regarding the number of calls made on a typical working day from each telephone of the sample centres are as summarized in Table 2.

Table 2 : Number of calls made from selected centres.

Centre	M	Calls made				Y_i	$\bar{Y}_i$
1	4	26	34	27	25	112	28
2	4	44	33	28	31	136	34
3	4	18	33	25	28	104	26
4	4	37	21	22	40	120	30
5	4	23	34	42	29	128	32

Estimate the average number of daily calls per telephone made from all the 25 centres. Also, estimate the relative efficiency of the estimator used with respect to the usual sample mean estimator, from the sample selected above.

Solution. Here, N=25, n=5 and M=4. The sample cluster means are given in the last column of Table 2. The estimate of average number of daily calls can be computed using estimates $\hat{\bar{Y}}_c$ for equal clusters. Using values from the last column of Table 2, we have

$$\hat{\bar{Y}}_c = \frac{1}{n} \sum_{i=1}^{n} \bar{Y}_i = \frac{1}{5}(28 + 34 + 26 + 30 + 30) = 30.$$

Also, since the variance estimator $\hat{V}(\hat{\bar{Y}}_c)$ for equal clusters is given by the relation

$$\hat{V}(\hat{\bar{Y}}_c) = \frac{N-n}{N\,n(n-1)} \left(\sum_{i=1}^{n} \bar{Y}_i^2 - n\,\hat{\bar{Y}}_c^2\right),$$

so, on making substitution, we get

$$\hat{V}(\hat{\bar{Y}}_c) = \frac{25-5}{(25)(5)(4)}\left[(28)^2 + (34)^2 + \ldots + (32)^2 - 5(30)^2\right] = 1.6.$$

Now, we know that the variance estimator of the sample mean estimator $\bar{Y}$ is given by the relation

$$\hat{V}(\hat{\bar{Y}}) = \frac{N-n}{(NM-1)n}\left[\frac{1}{nM}\sum_{i=1}^{n}\sum_{j=1}^{M} Y_{ij}^2 + \hat{V}(\hat{\bar{Y}}_c) - \hat{\bar{Y}}_c^2\right].$$

To make calculations easy, let us first compute the term involving sum of squares of all the individual observations. Once again, using values from above table, we get

$$\sum_{i=1}^{n}\sum_{j=1}^{M} Y_{ij}^2 = (26)^2 + (34)^2 + \ldots + (29)^2 = 18962.$$

Thus, by above stated relations,

$$\hat{V}(\hat{\bar{Y}}) = \frac{25-5}{[(25)(4)-1]5}\left[\frac{189621}{(5)(4)} + 1.6 - (30)^2\right] = 2.0081.$$

Finally, the estimate of percent relative efficiency will be

$$\hat{RE} = \frac{\hat{V}(\hat{\bar{Y}})}{\hat{V}(\hat{\bar{Y}}_c)} \times 100 = \frac{2.0081}{1.6}(100) = 125.5.$$

So, we can infer that here cluster sampling of centres is more efficient than the usual sample mean estimator when individual telephones would have been selected.

———— × ————

Now you try the following exercise.

E3) Suppose from a total of 120 bearing trees of guaua in a village, 5 clusters of 4 trees each are selected and (hypothetical) yield (in kg) recorded is as given in the following table.

Cluster	1st tree	2nd tree	3rd tree	4th tree
1	5	4	1	15
2	11	1	4	7
3	26	10	19	11
4	7	15	12	10
5	2	22	8	6

Estimate average yield (in kg) per tree of guaua along with its standard error. Also, estimate the relative efficiency of the estimator.

In this section, we discussed sampling procedures in which all the elements of the selected clusters were enumerated. This scheme, as you may have observed, is convenient and economical but the method restricts the spread of the sample over a population which generally reduces the efficiency of the estimator.

We now turn to the situation in which we first select the clusters and then randomly choose a specified number of units from clusters selected before. This procedure is known as **two-stage sampling** or **sub-sampling**.

14.3 MULTISTAGE SAMPLING

From previous section, recall the example of cluster sampling in which we grouped blocks of a city into clusters of approximately equal population. Suppose, instead of interviewing all households in sample clusters, we make a *random choice of households within each sample clusters*. That is, a sample is now selected in two stages – first

select a sample of **clusters, called** ***first-stage*** or **primary sampling units**, and then select a sample of elements within sample clusters.

We have the following three advantages of this sampling procedure.

(1) Lists have to be prepared for the selected *primary sampling units* and susequent stage units only;

(2) It is easy to check the correctness of the list; and

(3) A sample gets concentrated in the selected *primary sampling units* and this reduces costs of travel, etc.

In this course, we shall discuss two-stage sampling. In Table 3 below, we define notations that we shall need while discussing with you various relations used in the two-stage estimation procedure.

Table 3. Notations used in this section and their meanings.

N	number of *primary stage units* (**psu's**) in a population
n	number of **psu's** *selected in the sample*
M	number of *second stage units* (**ssu's**) in each **psu**
m	number of **ssu's** *selected* from M **ssu's**
$M_0 = N\,M$	total number of **ssu's** *in a population*
Y_{ij}	value of the study variable y for *jth* **ssu** of the *ith* **psu**, $j = 1, 2, \ldots, M; i = 1, 2, \ldots, N$
y_{ij}	value of the study variable for *jth* selected **ssu** of the *ith* selected **psu**, $j = 1, 2, \ldots, m; i = 1, 2, \ldots, n$
$Y_i = \sum_{j=1}^{M} Y_{ij}$	total of Y-values for the *ith* **psu**
$Y = \sum_{i=1}^{N} Y_i$	population total of Y-values
$\overline{Y}_i = \frac{Y_i}{M}$	population mean for *ith* **psu**
$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} \overline{Y}_i$	population mean for the study variable
$y_i = \sum_{j=1}^{m} y_{ij}$	sample total for the *ith* **psu**
$y = \sum_{i=1}^{n} y_i$	total of y-values for the whole sample
$\overline{y}_i = \frac{y_i}{m}$	sample mean for the *ith* **psu**

To familiarize yourself with these notations, try the following exercise.

E4) There are 50 fields in a village sown with wheat and each is divided into 8 plots of equal size. Out of 50 fields, 5 are selected by *without replacement simple random sampling* method. Again, from each selected field, 2 plots are chosen by *without replacement simple random sampling* method. The yield in kg/plot recorded is as given in the following table.

Selected field	Plot-1	Plot-2
1	4.16	4.76
2	5.40	3.52
3	4.12	3.73
4	4.38	5.67
5	5.31	2.59

Estimate the quantities $\overline{Y}$ and $\overline{y}_i$.

14.3.1 Preliminaries

As said above, in a *two-stage* sampling design, sample clusters form the units of sampling at the first stage, called *primary stage units* (**psu's**, in short). Then, the elements within clusters are called *second stage units* (**ssu's**, in short). It is now clear that this procedure can be generalised to three or more stages and that is why it is called **multi-stage sampling**.

For example, in a survey for estimating yield of a crop, a block in a district may be taken for *primary stage units*, villages within blocks as *second stage units*, the crop fields within village as *third-stage units*, and a plot of specified shape and size within field as the *ultimate unit* of sampling.

Multistage sampling has been found to be very useful in practice and is commonly used in large-scale surveys. This sampling procedure is a compromise between cluster sampling and direct sampling of units. Furthermore, this design is more flexible as it permits the use of different sample selection procedures at different stages.

It is important to mention here that multi-stage sampling is only choice in a number of practical situations, especially when a satisfactory sampling frame of ultimate-stage units is not readily available and cost of obtaining this information is large and time consuming.

In a *multi-stage sampling* procedure, the basic idea used in the estimation of population parameters is that of building up estimates from the bottom (last stage units) to the top. It is desirable that you keep this principle in mind while reading through the next part of the section.

14.3.2 Estimation of Population Mean

As we said before, only *two stage sampling* procedure will be discussed in this unit. Also, we shall assume that the first stage units are of equal size and that units at the first and second stage are selected by *without replacement simple random sampling* procedure.

Now to select a sample, we use a frame listing all the N *psu's* in the population. A *without replacement simple random sample* of n psu's can then be drawn using procedure as described in Unit 12. So, a frame listing all the M second stage units in i-th selected psu ($i = 1, 2, \ldots, n$) is obtained. Finally, a without replacement simple random sample of m units is drawn from the i-th selected psu, $1 \leq i \leq n$, containing M second stage units.

Throughout, the suffix *2s* refers to **two-stage sampling** procedure.

Then, the sample mean in this situation is given by

$$\hat{\bar{Y}}_{2s} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i.$$

This relation is based on all the nm units in the sample and is an unbiased estimator of population mean $\bar{Y}$.

For example, suppose in some locality of a city there are N *mohalla's* and we select one *mohalla* at random. Let the selected *mohalla* contains M households out of which m are selected at random. Now, we collect information on y (e.g., let it be weekly expenditure on fruits) from every household in the sample. Then, the sample mean $\bar{y}_1$ estimates the average (expenditure per household) in the *mohalla* and $M\bar{y}_1$ estimates the total (expenditure on fruits) for the whole *mohalla*. But, since this *mohalla* was selected at random from a total of N in the locality, the estimate of the total in the locality is $NM\bar{y}_1$.

Now suppose that not 1 but n *mohalla's* from N were selected without replacement with equal probabilities and each selected *mohalla* contain M household. Now, at the second stage of the sampling, we atke a random sample of m households from each *mohalla*.

Then, as before, $M\bar{y}_i$ $(i = 1, 2, \ldots, n)$ gives the estimate of the total (expenditure on fruits) for the ith *mohalla*. Therefore, $M\sum_{i=1}^{n} \bar{y}_i$ will give estimate for the total in the sample *mohalla's*. Hence,

$$\hat{\bar{Y}} = \frac{1}{nM}\left(M\sum_{i=1}^{n} \bar{y}_i\right) = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$$

provides the estimate for the total expenditure on fruits for the whole locality.

Again, the variance of the estimator $\hat{\bar{Y}}_{2s}$ is given by the relation

$$V(\hat{\bar{Y}}_{2s}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) S_w^2, \text{ where}$$

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\bar{Y}_i - \bar{Y})^2 \text{ and } S_w^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M} (Y_{ij} - \bar{Y}_i)^2$$

Furthermore, an unbiased estimator of $V(\hat{\bar{Y}}_{2s})$ is given by the relations

$$\hat{V}(\hat{\bar{Y}}_{2s}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{N}\left(\frac{1}{m} - \frac{1}{M}\right) s_w^2$$

$$= \left(\frac{N-n}{Nn}\right) s_b^2 + \frac{1}{N}\left(\frac{M-m}{NM}\right) s_w^2, \text{ where}$$

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(\bar{y}_i - \hat{\bar{Y}}_{2s}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} \bar{y}_i^2 - n\,\hat{\bar{Y}}_{2s}^2\right] \quad \text{and}$$

$$s_w^2 = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2$$

$$= \frac{1}{n(m-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 - m\sum_{i=1}^{m} \bar{y}_i^2\right]$$

Observe that in the expression for the variance of the sample mean $\hat{\bar{Y}}_{2s}$ there are two components wherein the first represents the contribution arising from sampling of first-stage units and the second is arising from subsampling within the selected first-stage units.

It is important to note the following two special cases of two stage sampling procedure.

(i) $n = N$, corresponds to stratified sampling with N first stage units as strata and m units drawn from each stratum; and

(ii) $m = M$, corresponds to cluster sampling.

Let us now consider a practical situation working of which will help us understand the use of above stated relations.

Problem 5. Assume that in Problem 4, from each selected centres, 2 telephones were chosen by *without replacement simple random sampling* method. The following table gives the data related to the number of calls made on a typical working day from chosen telephone of the selected centres.

Center	M	m	Calls made		y_i	$\bar{y}_i$
1	4	2	26	34	60	30
2	4	2	44	28	72	36
3	4	2	33	25	58	29
4	4	2	37	21	58	29
5	4	2	23	29	52	26

Estimate the average number of daily calls per telephone made from all the 25 centres alongwith its estimate of variance.

Solution. Here, $N = 25, n = 5, M = 4$, and $m = 2$. Also, observe that the sample means for selected cluster (psu) is as given in the last column of the table. Now, the estimate of average number of daily calls can be computed using estimator $\hat{\bar{Y}}_{2s}$. Thus,

using values from the last column of above table, we get

$$\hat{\bar{Y}}_{2s} = \frac{1}{5}\sum_{i=1}^{5}\bar{y}_i = \frac{1}{5}[30 + 36 + 29 + 29 + 26] = 30.$$

Also, we know that the estimated variance of $\hat{\bar{Y}}_{2s}$ is given by the relation

$$\hat{V}(\hat{\bar{Y}}_{2s}) = \left(\frac{1}{n} - \frac{1}{N}\right)s_b^2 + \frac{1}{N}\left(\frac{1}{m} - \frac{1}{M}\right)s_w^2.$$

Here, using values from table and the value of $\hat{\bar{Y}}_{2s}$, we get

$$\begin{aligned} s_b^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\bar{y}_i^2 - n\hat{\bar{Y}}_{2s}^2\right] \\ &= \frac{1}{5-1}[(30)^2 + (36)^2 + \ldots + (26)^2 - 5(30)^2] \\ &= \frac{1}{4}[4554 - 4500] = 13.5. \end{aligned}$$

Next, we calculate s_w^2. For, we first compute the term involving sum of squares of all the individual observations. Thus, using values from above table, we get

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2 &= (26)^2 + (34)^2 + \ldots + (29)^2 = 9446, \text{ and} \\ \sum_{i=1}^{n}\bar{y}_i^2 &= (30)^2 + (36)^2 + \ldots + (26)^2 = 4554. \end{aligned}$$

Thus,

$$\begin{aligned} s_w^2 &= \frac{1}{n(m-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2 - m\sum_{i=1}^{m}\bar{y}_i^2\right] \\ &= \frac{1}{5(2-1)}[9446 - 2 \times 4554] \\ &= \frac{1}{5}[338] = 67.6. \end{aligned}$$

Hence, the estimate of variance is

$$\begin{aligned} \hat{V}(\hat{\bar{Y}}_{2s}) &= \left(\frac{N-n}{Nn}\right)S_b^2 + \frac{1}{N}\left(\frac{M-m}{Mm}\right)S_w^2 \\ &= \left(\frac{25-5}{25 \times 5}\right)(13.5) + \frac{1}{25}\left(\frac{4-2}{4 \times 2}\right)(67.6) \\ &= \left(\frac{20}{125}\right)(13.5) + \left(\frac{2}{200}\right)(67.6) = 2.16 + 0.676 = 2.836. \end{aligned}$$

It is generally observed that in the variance expressions given above, the contributions due to first stage sampling is much larger than the contributions due to second stage. So, while estimating the variance, it may be approximated by the expression $\frac{s_b^2}{n}$.

A distinct advantage of multistage sampling is that the sampling variance may be broken up into as many components as there are stages. The expressions for unequal psu's are also available. However, we have already decided to consider only the case of equal psu's. The concept of multistage sampling is so common that it is difficult to visualise a real life survey situation where it has not been used.

Solve the following exercise.

E5) Assume that in E3, from each selected cluster, 2 trees were chosen by *without replacement simple random sampling* method. Generate the data for this situation and estimate the average yield (in kg) per tree of guaua of the village along with its standard error.

With this we have to the end of the unit. Let us summarise what we have discussed in this unit.

14.4 SUMMARY

In this unit we have discussed the following points.

1. A number of examples are given illustrating the basic principles of cluster and multistage sampling. Also, we talked about some advantages that these sampling techniques have.
2. For equal and unequal size of clusters, some of the relations used in estimating the population are discussed. A number of examples are discussed to illustrate their use. Also, relative efficiency (RE) aspect of cluster sampling is discussed with help of some examples.
3. Examples are dicussed to illustrate the use of the following formulations for estimating population mean in case of two-stage sampling method.

$$\text{(a)}\quad \hat{\bar{Y}}_{2s} = \frac{1}{n\,m}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i.$$

$$\text{(b)}\quad V(\hat{\bar{Y}}_{2s}) = \left(\frac{1}{n}-\frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right) S_w^2, \text{ where}$$

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\bar{Y}_i - \bar{Y})^2, \text{ and}$$

$$S_w^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M} (Y_{ij} - \bar{Y}_i)^2.$$

$$\text{(c)}\quad \hat{V}(\hat{\bar{Y}}_{2s}) = \left(\frac{1}{n}-\frac{1}{N}\right) s_b^2 + \frac{1}{N}\left(\frac{1}{m}-\frac{1}{M}\right) s_w^2$$

$$= \left(\frac{N-n}{N\,n}\right) s_b^2 + \frac{1}{N}\left(\frac{M-m}{m\,M}\right) s_w^2, \text{ where}$$

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(\bar{y}_i - \hat{\bar{Y}}_{2s}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} \bar{y}_i^2 - n\,\hat{\bar{Y}}_{2s}^2\right], \text{ and}$$

$$s_w^2 = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2$$

$$= \frac{1}{n(m-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2 - m\sum_{i=1}^{m} \bar{y}_i^2\right].$$

14.5 ANSWERS/SOLUTIONS

E1) Of course, this you will have to do it yourself.

E2) Change the figures and proceed as in Problem 3.

E3) (*Hint*) Proceed as in Problem 3 to find the values of $\hat{\bar{Y}}_c$ and $\hat{V}(\hat{\bar{Y}}_c)$. Finally, proceed as in Problem 4 to caculate $\hat{V}(\hat{\bar{Y}})$. Hence, these values will give

$$RE = \frac{\hat{V}(\hat{\bar{Y}})}{\hat{V}(\hat{\bar{Y}}_c)}.$$

E4) Do it yourself, using relations given in Table 3.

E5) Let the generated data be as in the following table.

Cluster	M	m	ssu's		y_i	$\bar{y}_i$
1	4	2	5	15	20	10
2	4	2	1	7	8	4
3	4	2	26	10	36	18
4	4	2	7	15	22	11
5	4	2	22	6	28	14

Here, $N = 120$, $n = 5, M = 4$, and $m = 2$. Also, the sample means for selected cluster (psu) are as given in the last column of the table. Now, the estimate of average number of daily calls can be computed using estimator $\hat{\bar{Y}}_{2s}$. Using values from the last column of above table, we get

$$\hat{\bar{Y}}_{2s} = \frac{1}{5}\sum_{i=1}^{5}\bar{y}_i = \frac{1}{5}[10 + 4 + 18 + 11 + 14] = 11.4.$$

Also, we know that the estimated variance of $\hat{\bar{Y}}_{2s}$ is given by the relation

$$\hat{V}(\hat{\bar{Y}}_{2s}) = \left(\frac{1}{n} - \frac{1}{N}\right)s_b^2 + \frac{1}{N}\left(\frac{1}{m} - \frac{1}{M}\right)s_w^2.$$

Here, using values from table and the value of $\hat{\bar{Y}}_{2s}$, we get

$$\begin{aligned} s_b^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\bar{y}_i^2 - n\hat{\bar{Y}}_{2s}^2\right] \\ &= \frac{1}{5-1}[(10)^2 + (4)^2 + \ldots + (14)^2 - 5(11.4)^2] \\ &= \frac{1}{4}[757 - 649.8] = 26.8. \end{aligned}$$

Next, to calculate s_w^2, we first compute the term involving sum of squares of all the individual observations. Thus, using values from above table, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2 = (26)^2 + (34)^2 + \ldots + (29)^2 = 9446, \text{ and}$$

$$\sum_{i=1}^{n}\bar{y}_i^2 = (30)^2 + (36)^2 + \ldots + (26)^2 = 4554.$$

Thus,

$$\begin{aligned} s_w^2 &= \frac{1}{n(m-1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}^2 - m\sum_{i=1}^{m}\bar{y}_i^2\right] \\ &= \frac{1}{5(2-1)}[9446 - 2 \times 4554] \\ &= \frac{1}{5}[338] = 67.6. \end{aligned}$$

Thus, the estimate of variance is

$$\begin{aligned} \hat{V}(\hat{\bar{Y}}_{2s}) &= \left(\frac{N-n}{Nn}\right)s_b^2 + \frac{1}{N}\left(\frac{M-m}{Mm}\right)s_w^2 \\ &= \left(\frac{25-5}{25 \times 5}\right)(26.8) + \frac{1}{25}\left(\frac{4-2}{4 \times 2}\right)(67.6) \\ &= \left(\frac{20}{125}\right)(26.8) + \left(\frac{2}{200}\right)(67.6) = 4.29 + 0.676 = 4.97. \end{aligned}$$