***CHAPTER 5***

# MULTISTAGE CLUSTER SAMPLING

(**With Equal Probabilities)**

**5.1 INTRODUCTION**

When the sample is selected from the population and information is recorded from the selected units then the sampling scheme is referred to as the single stage sampling. The simple random sampling and stratified random sampling schemes described before may be considered as single stage sampling schemes. Often it happens that the population is divided into larger units which are further divided into smaller units and initially the sample of larger unit is selected. From these selected larger units a sub-sample of smaller units is selected and studied. If the sampling is stop after these two stages then the sampling scheme is called a two stage sampling. If the number of stages increases then the scheme is called the Multistage sampling. Kendall and Buckland (\_\_\_\_) in Dictionary of Statistical Terms define the multistage sample as one “which is selected by stages, the sample units at each stage being sub-sampled from the (larger) units chosen at the previous stage”. In other words multistage sampling is the term applied to that kind of sampling in which selection is carried out in stages where the sampling units at each stage are sub-sampled from the larger units chosen at the previous stage. The practical example of multistage sampling can be described in the situation where a municipality is divided into certain number of zones and a number of zones can be randomly selected as **first-stage** sampling units. Within each selected zone a number of schools can be chosen at random as **second-stage** sampling units. Within each school drawn in the sample, a sample of students can be randomly selected as **third-stage** sampling units. This is an example of a three-stage sampling, where zone is **first-stage unit**, school is **second-stage unit** and students selected are the **third-stage units.** In multistage sampling the units selected at the last stage are studied. In the present example students drawn within schools composed the sample to be analyzed. This sort of sampling scheme is often used in combination with area sampling and cluster sampling. In-fact at the first stage the entire population is divided into primary stage units (P.S.U’s) and a selection is made from them. At each successive stage, smaller sampling units are defined within the units selected at the previous stage, and a further selection is made within each of them.

Typical instances of multistage sampling are the sample survey conducted by Agriculture Census Organization in Pakistan. The **primary sampling units** (P.S.U’s) are the Tehsil (administrative unit). The **secondary sampling units** (S.S.U’s) are villages within each tehsil. The **tertiary sampling units (third stage)**units (T.S.U’s) are fields and **quaternary sampling units(fourth stage units)** (Q.S.U’s) are the plots. Multistage sampling is most frequently used in field surveys were the ultimate final stage units of selection are scattered geographically and where the maintenance of complete lists is difficult. In such situations clusters are formed on the area basis. Two principal advantages arise from this arrangement. First is that the listing is greatly simplified, as at each stage of selection it is necessary to define and prepare lists of only those units contained within the sample units selected at the previous stage. Secondly, the field operation involved in the collection of survey data can be restricted to relatively small number of compact areas, with the result that the travel involved appreciably reduced.

In the design of multistage samples, considerable choice is usually possible in deciding the number of sampling stages, in the definition of sampling units and in choosing the number of units to be selected at each stage. Further more, choice may be made at each stage as to whether selection is to be made with equal or unequal probabilities, and (for instance) whether it is to be with or without replacement. Advantage may be taken of this flexibility of design to achieve any required intensity of clustering within areas, with the objective of reducing the cost of survey operations to some specified level, and minimizing the sample error for that given cost. Even cost of selection of sample is much low as compared to the selection of simple random sampling or systematic sampling. Indeed, selection of a simple random sample of houses in a city of the size of Karachi would be prohibitively expensive. The simplest form of multistage sampling is two stage sampling. In-fact multistage sampling has many advantages over other (i.e. simple random sampling, systematic sampling, single stage cluster sampling) sampling schemes. Some of these are listed below:

1. In developing areas where a suitable and up-to-date frame is not available, it becomes necessity to construct a rough frame for some larger areas. Some of the larger areas are selected and their maps are prepared, frames of selected areas are constructed and sample is drawn from these frames.
2. It is much cheaper and quicker than any other sampling design.
3. While collecting information on units at every stage for construction of frame the ancillary information can be used to improve the estimates and efficiency of the sample designs.
4. In large scale surveys where complete and up-to-date reliable frames are not available it may still be convenient to adopt this scheme in order to avoid high cost, due to travel, location of units, etc. and reduce the supervision work on the scattered areas, i.e. selection of simple random sample of dwelling in a city of the size of Karachi would be very much expensive.
5. Multistage scheme warrant a strict vigilance over interviewer and may increase efficient organizational setup at each level of responses.
6. The scheme may be designed such that it increases the intra-class correlation, thereby improving efficiencies of sample estimates.

Multistage sampling occupies a central role both in theory and in the applications of unequal probability sampling. It was in the context of multistage sampling that unequal probability sampling was first suggested by Hansen and Hurwitz (1943). Multistage unequal probability sampling is often used in area surveys or surveys of individuals and households. Here multistage sampling is used partly to overcome the problem that lists of the ultimate sampling units are typically not available, and partly to reduce travel costs by ensuring that the sample units are geographically clustered. Unequal probability sampling, in this context is used to reduce sampling errors.

In this chapter we focus on multistage sampling with equal probabilities only. The problem of unequal probabilities in multistage sampling will be discussed in Chapter 12.

We now give some basic notations for multistage sampling:

**5.2 NOTATION**

|  |  |  |
| --- | --- | --- |
| *N* | = | The number of first stage units in the population. |
| *Mi*  | = | The number of second stage units in the *i*th first stage unit. |
| *Qij* | = | The number of third stage units in the *i*th first stage and *j*th second stage units. |
| *n* | = | The number of first stage units in the sample. |
| *mi* | = | The number of second stage units to be selected from the *i*th first stage sample unit. |
|  |  |  |
| *qij* | = | The number of third stage units to be selected *i*th first and *j*th second stage sample unit.  |
|  |  |  |
| *Yi* | = | Population values for the *i*th first stage unit. |
| *Yij* | = | Population value for the *j*th second stage units the *i*th first stage. |
| *Yijk* | = | Population value for the *k*th third stage unit with the *j*th second stage and *i*th first stage unit. |

*yi*, *yij* and *yijk* will be used for the sample values corresponding to *Yi*, *Yij* and *Yijk*.

*Zi*, *Zij* and *Zijk* and *zi*, *zij* and *zijk* will be used as bench mark variable corresponding to *Yi*, *Yij* and *Yijk* and *yi*, *yij* and *yijk* respectively.

**5.3 VARIANCE ESTIMATION FOR MULTISTAGE SAMPLING**

A basic principle of multistage sampling is that when selection and estimation take place independently at various stages the variance of an unbiased estimator that arises from each of these stages should be added. In particular the total variance of such a multistage estimator is equal to the variance arising from the first stage plus that arising from subsequent stages. For a two stage sample the variance may formerly be written as:

  (5.3.1)

where *E*1 denotes the expectation and *V*1 is the variance over all first stage samples, *E*2 denotes the expectation over all second stage samples, and  is the conditional variance of y' subject to the selection of a particular first stage sample. An obvious extension of above expression for three stages sample is:

   (5.3.2)

and for k stage sampling, the expression is:

 

  (5.3.3)

If the totals for the first stage sample units are known exactly it would be possible to estimate , the first stage variance, in exactly the same fashion as in single stage sampling and same is true for other stages. The basic problem for multistage variance estimation is that there totals have to be replaced by estimates from the lower stages of sampling, and that this introduces a component from these lower stage which in general bears no direct relationship to the actual variance from these lower stages.

**5.4 TWO STAGE SAMPLING**

The simplest of the multistage sampling is when the sampling procedure is stopped after only two stages. The sampling in this case is called the Two Stage sampling. In this section we have derived the expression for variance of the estimate of population total when a two stage sample is available.

The main idea is illustrated below:

Let a population of *N* primary stage units (clusters) of sizes *Mi* (*i* = 1, 2, 3, …, *N*) is given. We may select a sample of *n* primary stage units (clusters and within each selected primary stage unit (cluster), a sample of size *mi* secondary stage units is drawn. If a sample of *n* clusters have been selected from the population of *N* clusters and from each selected cluster a sample of size *mi* has been selected from *Mi* element then an unbiased estimator of population total *Y* is given as:

   (5.4.1)

  as 

 , as ,

where *yij* is the value of *j*th second stage unit from the *i*th primary stage unit.

The unbiasedness of the estimate is proved in the following theorem:

#### THEOREM 5.1

In two stage sampling estimated population total is an unbiased estimator of *Y*.

**PROOF:**

Consider 

Now 

 

Taking first stage expectation 

 .◊

The variance of estimator  is derived in the following theorem.

#### THEOREM 5.2

In two stage sampling with equal probabilities at both stages, the sampling variance of  is:

  (5.4.2)

where .

#### PROOF:

We know that

 

  (5.4.3)

After substituting the value of  from (5.4.1) in (5.4.3) and squaring we get

 

  (5.4.4)

where 

Taking the third term and applying second stage expectation, we have:

  as 

So (5.4.4) is:

  (5.4.5)

Now , (Primary stage variance) (5.4.6)

Also 

as 

So 

  (5.4.7)

The second term is zero as 

So we left with

 

= (5.4.8)

 

 

 

 

 

  (5.4.9)

where 

Putting (5.4.6) and (5.4.9) in (5.4.5) we get (5.4.2) ◇

Alternatively (5.4.2) may be written as:

 , (5.4.10)

where  and 

If  then all the clusters are equal in size and sampling fraction in each cluster at the second stage in uniform then (5.4.10) is

 , (5.4.11)

where .

Note that of two stage sampling is sum of two components. The first component is variation of primary stage units (between variance) and second component comes from the variance of second stage units within primary stage unit (within variance).

The variance for sample mean  may be written from (5.4.10) as

 . (5.4.12)

If all the primary stage units are of equal size then variance of sample mean is

 . (5.4.13)

 may alternatively be derived by using (5.3.1). For this we have

 

  (5.4.14)

and

 

 (5.4.15)

Substituting (5.4.14) and (5.4.15) in (5.3.1) we obtain (5.4.2). ◇

**5.4.1 Unbiased Variance Estimator**

#### The unbiased variance estimator for the estimator of two stage sampling has been derived in the following theorem.

#### THEOREM 5.3

An unbiased variance estimator of Var(y') is

 , (5.4.16)

where ,

and .

#### PROOF

Starting with the expectation of  we have

  

  (5.4.17)

Taking the first term of (5.4.17) we have:

  

 .

Applying the first stage expectation

  (5.4.18)

and  

  (5.4.19)

Since the second stage covariance between  and  is zero, as second stage selection is independent.

Taking the first stage expectation of (5.4.19)

 

  (5.4.20)

Substituting (5.4.18) and (5.4.20) in (5.4.17)

 

 (5.4.21)

Since in the first term of (5.4.2) is a multiple factor of  so we also multiply (5.4.21) by the same factor and on simplification

 

or 

since 

Therefore 

is an unbiased estimator of (5.4.2). ◇

If all the clusters are equal in size and same number of units are selected at the second stage then

  (5.4.22)

Expression (5.4.22) provides the unbiased variance estimator for estimator of population total in two stage sampling with equal probabilities.

Some further aspects of multistage sampling with arbitrary probabilities will be discussed in chapter 12 alongside the exercises on the topic.