***CHAPTER 4***

**SINGLE STAGE CLUSTER SAMPLING**

**2.1 INTRODUCTION**

Historically speaking the term **cluster** wasfirstused by Hansen and Hurwitz (1942) to describe a group of elements that constitute a sampling unit, though the technique of cluster sampling had at that time been used for over a century [Ström (1830)]. Cluster sampling is a procedure in which population units are divided into convenient number of groups, called **clusters;** each **cluster** containing some elements; a random sampling of some **cluster** is selected and each selected cluster is studied in full. If all the elements in the sampled cluster are examined in full, it is known as **single stage cluster sampling**. Sometimes cluster are known as **primary units** in the context of multistage sampling and elements within each cluster are called **secondary units**. In a survey for estimating the wheat area in a certain tehsil (administrative units), tehsil is divided into villages, the sample villages is selected, each selected village is completely studied and population estimates are made from these information. In this problem villages are known as clusters (primary nits) and farms within the villages are called elements (secondary units).

The concept of cluster sampling was developed for the cases where the list of the elements is not available. For example, in a population survey, a list of households is available where as a list of the persons is not. Similarly in an agriculture survey, list of the villages is available whereas list of the farmers is not available. When such situation arises cluster sampling is more useful then other sampling designs.

Cluster sampling consists of group of elements, approach to the elements is faster, easier and more convenient than other sampling procedures. For example, in a population survey it is faster and convenient to collect information in a cluster than from the sample of same number of households directly selected with simple random sampling procedure. In an agriculture survey it is also convenient, faster and easier to collect information from farmers in a cluster than the same number of farmers selected with simple random sampling.

Cost, which is a major factor in a survey, will be less if the elements are grouped in a cluster rather than randomly dispersed throughout the area.

Since cluster sampling is not a true representative sampling method as compared to simple random sampling procedure, the efficiency of this method is generally less. The efficiency of clustering sampling depends on the size of the cluster and nature of units within cluster. If the size is large the efficiency is decreased, if the size is small and number of clusters are more the efficiency is increased. In a household survey, the households; if selected randomly and independently will represent the whole population and thereby provide better estimate then cluster sampling. Experiences have shown that cluster sampling procedure is more precise than simple random procedure if the variation within the cluster is more than overall variation. Hence clusters should be formed in such a way that individual (elements) within the cluster vary as much as possible so that maximum precision can be obtained.

Cluster sampling procedure is different from stratified sampling in the sense that in the former case elements within the groups vary as much as possible whereas in latter case individual within the groups are as much homogeneous as possible. In cluster sampling simple random sampling of n cluster is selected from a population of N clusters and each selected cluster is studied in full, whereas in the stratified sampling simple random sample is selected within each group and selection is independent within the groups. Cluster should be formed in such a way that the variances within the cluster is high whereas in stratified sampling variances within the strata is low.

**4.2 CLUSTER SAMPLING WITH CLUSTERS   
OF EQUAL SIZES.**

In this section we give the cluster sampling with equal number of units within each cluster. The illustration of this sort of sampling can be given as under:

Suppose a population is divided into N clusters, each containing M elements as:-

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C l u s t e r s | | | | | | | | |
| Units ↓j i→ | 1 | 2 | 3 | ......... | I | ......... | *N* |  |
| 1 |  |  |  | ......... |  | ......... |  |  |
| 2 |  |  |  | ......... |  | ......... |  |  |
| 3 |  |  |  | ......... |  | ......... |  |  |
| . | . | . | . | ......... | . | ......... | . |  |
| . | . | . | . | ......... | . | ......... | . |  |
| . | . | . | . | ......... | . | ......... | . |  |
| *j* |  |  |  | ......... |  | ......... |  |  |
| . | . | . | . | ......... | . | ......... | . |  |
| . | . | . | . | ......... | . | ......... | . |  |
| . | . | . | . | ......... | . | ......... | . |  |
| *M* |  |  |  | ......... |  | ......... |  |  |
| Mean | ‾ | ‾ | ‾ | ......... | ‾ | ......... | ‾ |  |

where  denotes the *j*th element of the *i*th cluster (*i* = 1, 2, 3, ......... *N*)

(*j* = 1, 2, 3, ......... *M*)

Following notations are defined here:

Population mean 

Mean of the *i*th cluster 

Between cluster variance 

Within cluster variance 

Total variance  (4.2.1)

Total variance can be partitioned into between and within variance as:





 (4.2.2)

or

 (4.2.3)

These may be placed in the form of analysis of variance table as:-

|  |  |  |  |
| --- | --- | --- | --- |
| Source of variance | d.f. | S.S. | M.S.S. |
| Between-cluster  Variance | *N*–1 |  |  |
| Within-cluster  variance | *N*(*M*–1) |  |  |
| Total Variance | *NM*–1 |  |  |

If a sample of *n* clusters is drawn at random from a population of *N* clusters then

Sample mean 

Mean of the *i*th cluster 



 (4.2.5)

Total variance may be written analogous to (4.2.2) as

 (4.2.6)

If *n* and *M* are sufficiently large then (4.2.5) becomes:

 (4.2.7)

This may also be presented in the form of analysis of variance table

|  |  |  |  |
| --- | --- | --- | --- |
| Source of variance | d.f. | s.s. | M.s.s |
| Between-cluster  variance | *n*–1 |  |  |
| Within-cluster  variance | *n*(*M*–1) |  |  |
| Total Variance | *nM*–1 |  |  |

**4.3 EXPECTATION**

**THEOREM (4.1)**

In a simple random sample of size *n* clusters drawn without replacement from a population of *N* clusters of equal sizes, the sample mean  is an unbiased estimator of population mean .

# Proof

We know that

 (4.3.1)

Since all the clusters are equal in size  is clearly an unbiased estimator of population mean  (simple random sampling)

Taking expectation of (4.3.1) we get



 (4.3.2)

Note that if a simple random sample of *n* clusters is drawn from a population of *N* clusters, the estimated total  is an unbiased estimator of population total *Y* i.e.

 (4.3.3)

**4.4 VARIANCE AND UNBIASED VARIANCE ESTIMATOR**

The efficiency of cluster sample is decided on the basis of variance of estimator used in cluster sample. In this section we have derived the variance and variance estimator of mean of cluster sample. The variance and variance estimator is given in the following theorem.

**theorem (4.2)**

A simple random sample of n clusters is drawn without replacement from a population of N clusters, the variance of sample mean,  is

 (4.4.1)

# PROOF

Since random sample of n cluster is drawn from a population of *N* cluster and  are the means of n clusters, the variance of  is



Using theorem (2.2) from simple random sampling we get (4.4.1)

This may be put as



 (4.4.2)

The relative efficiency of  and  i.e. cluster sampling and simple random sampling is



i.e. the efficiency of cluster sampling increases as the  decreases. This may be shown using (4.2.2) with re-arrangement as

 (4.4.3)

This shows that as  (within variation) increase the efficiency of cluster sampling also increases.

 (4.4.4)

The estimator of population mean can also be based upon a sample of *nM* units drawn directly from a population of *NM* units. The variance of estimator, say , will differ from the estimator based on the cluster sample. This variance   
is given in the following theorem.

**THEOREM (4.3)**

A without replacement random sample of *nM* elements is drawn directly from a population of *NM* elements, instead of using clusters, then variance of sample mean  is

, (4.4.5)

# 

# PROOF

Since a random simple is drawn from *NM* elements, by Theorem (2.2) the variance may be written directly as

 (4.4.6)

 (4.4.7)

The variance of estimated total  is

 (4.4.8)

When population is grouped in clusters then the sample can be selected in two ways as illustrated earlier. The selection of sample is carried out with a view to decrease variance of the estimator. Generally speaking when population is clustered then sample of clusters is preferred over a simple random sample to save time and cost. Further, the sample of clusters prove more efficient then a simple random sample under certain condition. This condition is given in the following theorem.

**THEOREM (4.4)**

A simple random sample without replacement of *n* cluster is drawn from a population of *N* clusters of equal sizes and also same number of elements are drawn from a population of *NM* elements, the cluster sample will be more precise then simple random sample provided within cluster variance is more than over all variance.

# PROOF

Using (4.4.1) and (4.4.6) we have:

 (4.4.9)



(4.4.10)

From (4.2.2) we get

 (4.4.11)

Using (4.4.11) in (4.4.10) we get

 (4.4.12)

This shows that  provided .

Hence cluster sampling is more precise than simple random sampling with the same sample size if the units within the clusters vary more on the average than the units in the population as a whole. We can say greater the variation within the cluster the greater the precision of cluster sampling. This result is identical to systematic sampling but opposite to stratified random sampling.

**4.4.1 Variance in Terms of Intraclass Correlation Coefficient**

In numerous cases, a general measure is required which gives some indication to the accuracy of cluster sampling and to the effect of cluster size has on the sampling scheme. Many sampling statisticians attempted to derive some relationship, i.e. Smith (1938), Jessen (1942), Mahalanobis (1940), Cochran (1942), derived an algebric solution to the problem. However, Hansen and Hurwitz (1942) presented examples based on household samples which showed that the relationship previously assumed did not hold. They put forward a model in terms of the interclass correlation coefficient which becomes the appropriate concept in discussions on cluster sampling.

In general terms, the intraclass correlation coefficient ρ (measure of homogeneity), is simply the correlation between pairs which are in the sample.

 (4.4.13)

[In the case of cluster sampling with equal clusters, the number of terms in the numerator is *NM*(*M*–1) and in the denominator is (*NM*–1)].

The intraclass correlation can be used to give and alternative expression for variance of estimator in cluster sampling. For this we first note that:



 (4.4.14)

Using (4.2.1) and (4.4.13) in (4.4.14), we have



or

 (4.4.15)

**Theorem (4.5)**

If a simple random sample of *n* cluster is drawn from a finite population of *N* clusters of equal size the variance of sample mean  is

 (4.4.16)

# 

# PROOF

We know that

 (4.4.1)

Using (4.4.15) in (4.4.1), we get

 (4.4.17)

If *M* and *N* are large than (4.4.17) can be written as:

 (4.4.18)

Ignoring f.p.c. (4.4.18) becomes

 (4.4.19)

Variance formula for the estimation of total may be written as

 (4.4.20)

From (4.4.20) it can be seen that  not only depends on samples size *n*, but it also depends upon size of the cluster *M*, and intra-class correlation coefficient. If ρ is negative than  and in consequence the efficiency of cluster sampling is more. If ρ = 0 then  and if ρ = 1, (complete homogeneity) than.

The efficiency of cluster sampling can also be compared with simple random sampling. For this comparing (2.4.1) and (4.4.18), we get

 (4.4.21)

The relative efficiency of cluster sampling and random sampling of *nM* elements selected directly from *NM* elements can also be compared using (4.4.5) and (4.4.19), we get

 (5.4.22)

From (4.4.21) and (4.4.22) we can readily see that if M = 1 or P = 0 then

.

**4.4.2 Calculation of ρ.**

Experience shows that the magnitude of coefficient of intraclass correlation usually decreases as size of the cluster increases with rate of latter being faster. The knowledge of effect of ρ on the efficiency of cluster sampling is important so that full use can be made of the advantage of cluster sampling.

The simple ways for calculation of ρ are as follows:

i) From (4.4.14) we know that:

 (4.4.15)

We know that



Using this in (4.4.15) we have:

 (4.2.23)

or  (4.4.24)

or  (4.4.25)

ii) Another expression for P may be obtained by using (4.2.2) as under:

We have

 (4.2.2)

Substituting the value of  from (4.4.23) in (4.2.2) and on simplification we get

 (4.2.24)

or  (4.4.25)

**4.2.2 Estimated Variance**

In this section we have given the variance estimator of estimate of cluster sampling.

An unbiased estimator of (4.4.1) is:

 (4.4.28)

The variance of estimated total  is

 (4.4.29)

where

. (4.4.30)

An unbiased variance estimator of  is

 (4.4.31)

 (4.4.32)

Infact the value of ρ varies very considerably from population to population and from variable to variable within a given population but always lies in the range  usually ρ is positive clustering increase the variance of sample estimator.

It is instructive to consider what the implications are of various types of clustering on ρ:

1. When the units within each cluster are identical, but those in one cluster differ from those in another, then ρ = 1 and clustering will increase the variance to such an extent that it is equal to that obtained from a sample of *nM* units selected from a population of *NM*.
2. When the units within a cluster differ from each other, but each cluster is identical with every other cluster, then ρ = – 1/(*M*–1) and the variation is zero. This makes sense, in that a sample of only one cluster will supply perfect information about the entire population.
3. When the clusters are random aggregations of population units, then ρ = 0 and the variance of a cluster sample of *n* units is identical with the variance of a simple random sample of n units.

# EXAMPLE 4.1

A simple random sample of 15 cluster of 4 trees each, was selected out of 420 hearing trees of oranges to study the yield f oranges. The yield is recorded in (kg/trees) and is given as:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | **1st Tree** | **2nd Tree** | **3rd Tree** | **4th Tree** | **T** |
| 1 | 5.17 | 26.11 | 29.54 | 19.64 | 80.46 |
| 2 | 3.53 | 40.76 | 5.15 | 1.25 | 50.69 |
| 3 | 14.23 | 16.89 | 28.93 | 21.70 | 81.75 |
| 4 | 7.13 | 34.35 | 12.18 | 9.86 | 63.52 |
| 5 | 27.59 | 38.10 | 24.74 | 6.77 | 97.20 |
| 6 | 5.58 | 4.84 | 0.69 | 15.69 | 26.80 |
| 7 | 12.66 | 32.53 | 16.92 | 37.02 | 99.13 |
| 8 | 6.40 | 11.68 | 40.05 | 5.12 | 63.25 |
| 9 | 54.21 | 34.63 | 52.55 | 37.96 | 179.35 |
| 10 | 37.94 | 47.07 | 19.64 | 29.11 | 133.76 |
| 11 | 45.98 | 5.17 | 1.47 | 6.53 | 59.15 |
| 12 | 0.87 | 3.56 | 4.86 | 23.56 | 32.85 |
| 13 | 26.11 | 10.93 | 10.08 | 11.18 | 58.30 |
| 14 | 1.94 | 35.97 | 29.54 | 25.28 | 92.73 |
| 15 | 11.08 | 0.65 | 4.21 | 7.56 | 23.50 |

1. Estimate the average yield per tree as well as the production of oranges in the village along with standard error of your estimate.
2. Estimate the intra-cluster correlation coefficient between trees within clusters.
3. Estimate the efficiency of cluster sampling as compared to simple random sampling.

# SOLUTION

There are 420 trees and each cluster has 4 trees.

Therefore total number of clusters are . Also a sample of 15 cluster is selected.

Necessary calculations are given below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster |  |  |  |  |  |  |  |
| 1 | 20.1150 | 80.4600 | 1618.453 | 1966.802 | 116.116 | 440.6132 | 1.162 |
| 2 | 12.6725 | 50.6900 | 637.815 | 1700.685 | 354.290 | 159.4543 | 41.082 |
| 3 | 20.4375 | 81.7500 | 1670.766 | 1795.600 | 41.611 | 417.6914 | 1.961 |
| 4 | 15.8800 | 63.5200 | 1008.698 | 1476.332 | 155.878 | 252.1744 | 9.967 |
| 5 | 24.3000 | 97.2000 | 2361.960 | 2870.718 | 169.586 | 590.4900 | 27.699 |
| 6 | 6.7000 | 26.8000 | 179.560 | 301.214 | 40.551 | 44.8900 | 152.202 |
| 7 | 24.7825 | 99.1300 | 2456.689 | 2875.243 | 139.518 | 614.1723 | 33.011 |
| 8 | 15.8125 | 63.2500 | 1000.141 | 1807.599 | 269.153 | 250.0352 | 10.397 |
| 9 | 44.8375 | 179.3500 | 8041.606 | 8340.425 | 99.607 | 2010.4010 | 665.666 |
| 10 | 33.4400 | 133.7600 | 4472.934 | 4888.150 | 138.405 | 1118.2340 | 207.446 |
| 11 | 14.7875 | 59.1500 | 874.681 | 2185.150 | 437.004 | 218.6702 | 18.058 |
| 12 | 8.2125 | 32.8500 | 269.781 | 592.124 | 107.448 | 67.4452 | 117.170 |
| 13 | 14.5750 | 58.3000 | 849.723 | 1027.796 | 59.358 | 212.4316 | 19.909 |
| 14 | 23.1825 | 92.7300 | 2149.713 | 1027.796 | 219.860 | 537.4283 | 17.185 |
| 15 | 5.8750 | 23.5000 | 138.063 | 2809.294 | 20.001 | 34.5166 | 173.238 |
|  | 285.6100 |  | 27730.58 | 198.067 | 2368.387 | 6932.645 | 1496.154 |



Correction factor

B.S.S (Between clusters)

= 



T.S.S 



W.S.S (within clusters) 

These calculations may be tabulated as:

|  |  |  |  |
| --- | --- | --- | --- |
| *S.V* | *d.f* | *S.S* | *M.S.S* |
| B.S.S | n – 1 = 14 | 5982.243 | 427.303 |
| W.S.S | n (M – 1) = 45 | 7101.95 | 157.821 |
| T.S.S | n M – 1 = 59 | 13084.16 |  |

i) 

ii) 

iii) 



iv) , 





Interclass conclusion coefficient may be calculated as



or 

# 

# EXAMPLE 4.2

From an artificial population of 200 clusters of Mohallahs of equal size of 5 households a simple random sample without replacement of 15 clusters is selected and number of persons of each household are recorded as:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | **Number of persons in the households *yij*** | | | | |
| 1 | 5 | 6 | 3 | 5 | 3 |
| 2 | 5 | 8 | 3 | 8 | 3 |
| 3 | 3 | 2 | 5 | 6 | 2 |
| 4 | 5 | 3 | 2 | 2 | 6 |
| 5 | 7 | 3 | 5 | 7 | 8 |
| 6 | 7 | 3 | 7 | 2 | 5 |
| 7 | 8 | 3 | 8 | 8 | 4 |
| 8 | 2 | 5 | 3 | 2 | 2 |
| 9 | 3 | 7 | 3 | 5 | 6 |
| 10 | 3 | 4 | 3 | 5 | 6 |
| 11 | 6 | 7 | 3 | 2 | 3 |
| 12 | 3 | 7 | 5 | 8 | 6 |
| 13 | 7 | 8 | 3 | 5 | 9 |
| 14 | 5 | 6 | 8 | 5 | 6 |
| 15 | 7 | 3 | 8 | 3 | 3 |

Estimate the total number of persons in the 200 clusters and find the standard error of this total.

# SOLUTION

Necessary calculations are given below:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 22 | 27 | 18 | 18 | 30 | 24 | 31 | 14 | 24 | 21 | 21 | 29 | 32 | 30 | 24 |
| Average | 4.4 | 5.4 | 3.6 | 3.6 | 6.0 | 4.8 | 6.2 | 2.8 | 4.8 | 4.2 | 4.2 | 5.8 | 6.4 | 6.0 | 4.8 |

N = 200 n = 15 

  



The total number of persons in the population can be estimated as



This may be put in the form of analysis of variance table as

|  |  |  |  |
| --- | --- | --- | --- |
| Source of Variation | d.f. | Sum of Squares | M.S.S. |
| Between cluster | 14 | 82.267 | 5.876 |
| Within cluster | 60 | 234.400 | 3.907 |
| Total | 74 | 316.667 | 4.279 |

The variance of  may calculated as



Substituting the value from the table we have:



The variance may also be calculated using intra-class correlation coefficient and ρ is calculated as





**4.5 DETERMINATION OF CLUSTER SIZE**

It has been already pointed out that variance of sample mean in cluster sampling depends on the number of clusters, *S*2, and intra-class correlation coefficient ρ. We have also seen that ρ depends on the size of the cluster that varies from population to population. In this section the optimum level of cluster has been determined. Mahalanobis (1940, 42, 44), Hansen-Hurwitz (1942), Cochran (1942), H.M.F Smith (1938) and Sukhatme (1947) have considered in detail the question of optimum level of cluster for various field surveys and also in crop cutting experiments. Brewer (1977) has considered this problem for survey of farmers in the Indus Basin, Pakistan. Hanif and Ahmad (1977) have also determined the optimum level of cluster size for multipurpose sample surveys in Libya.

The problem is to find optimum value of cluster when the cost is involved.

Let us take a simple cost function.

 (4.5.1)

*C* = total cost, *C*0 = over head cost

*C*1 = cost per cluster *C*2 = cost per element in each cluster

From (4.4.18) the variance of cluster sample is:

 (4.4.18)

The objective is to find *M* by minimizing  subject to cost given in (4.5.1). The value of *n* from (4.5.1) is

 (4.5.2)

Substituting the value of *n* from (4.5.2) in (4.4.18) the function is



Partially differentiating with respect to *M* and equating to zero we have:



On simplification we get

 (4.5.3)

For different values of ρ, C1 and C2 the optimum value of M can be easily obtained.

For ρ = 0; *M* → ∞, for ρ = 1; M → 0

For ρ = 0.001;  and for ρ = 0.1; 

We see that as ρ increase *M* (size of cluster) decreases.

Another cost function given by Jessen (1942) is

 (4.5.4)

where

*C* = total cost,

*C*1 = cost of interview and cost of travel from one element   
 to another element, and

*C*2 = cost of travel between clusters,

The variance of is:

 (4.4.18)

differentiating with respect to *M* we have:

 (4.5.5)

also differentiating (4.5.4) with respect to *M* and on simplification

 (4.5.6)

Substituting *dn*/*dM* in (4.5.5) and on simplification, we get:

 (4.5.7)

Solving (4.5.4) as a quadratic in 

 (4.5.8)

From (4.5.7) and (4.5.8) we have



or



or

 (4.5.9)

Jessean (1942), Mahalanobis (1944), Hendricks (1944), Hansen, Hurwitz and Madow (1953), Cochran (1977) and Murthy (1968) have given an excellent discussion on this topic. They also used linear model in order to determine the size of the cluster. [For details see Cochran (1977)].

# EXAMPLE 4.3

For a survey on birth rate, given that the total cost, neglecting the over heads cost, is fixed at Rs, 20,000 and the enumerator cost per month is Rs. 300, what is the optimal size of sample if it is decided to select a cluster of persons, assuming that an enumerator has to spend, on average, two days in contacting the clusters and in other primary work, that he can enumerate an average 40 persons per day, the intraclass correlation coefficient is estimated at 0.001. (Som 1973).

# SOLUTION

Total cost = *C* = Rs. 20,000, Enumerator cost per month = Rs. 300

1 – man-day = 300/30 = Rs. 10, *C*1 = 2 man-day = Rs. 20

*C*2 = 1/40 man-day = 0.25, *C*1/*C*2 = 80, ρ = 0.001

The optimum size of cluster from (4.5.3) is

M = 

Total cost (neglecting overhead)

*C* = *C*1*n* + *nMC*2

20,000 = 20 *n* + *n*(283) (0.25) = *n* [20 + 283 x 0.25] = 90.75 *n*

The sample size is

*n* =  = 220 sample cluster

The total sample size is

*nM* = 283 x 220 = 62260

Optimum size of the cluster for typical ratio of *C*1/*C*2 and of ρ are given as:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0.001 | 0.002 | 0.003 | 0.004 | 0.005 |
| 50 | 224 | 158 | 129 | 112 | 100 |
| 75 | 274 | 194 | 158 | 137 | 122 |
| 100 | 316 | 224 | 183 | 158 | 141 |
| 125 | 373 | 264 | 215 | 187 | 158 |
| 150 | 287 | 274 | 223 | 194 | 173 |

# EXAMPLE 4

In a rural survey in which sampling unit is a cluster of *M* farms, the cost of taking a sample of n units is C = 4 t Mn + 60 , where t is the time in house spent getting the answers from a single farmer. If Rs. 20,000 is spent on the survey the find the value of *n* for *M* = 1, 10 and t = ½, 2 hours.

# SOLUTION

Since *C* = 2000, *M* = 1, 10 *T* = ½, 2

Cost function is *C* = 4 *t M n* + 60 

For *M* = 1, *t* =  we have

2000 = 4 .  . 1 n + 60 

or

*n* = 30  = 1000

* 
* *n*2 – 2900*n* + (1000)2 = 0
* *n* = 

*n* = 400, 2500.

For *M* = 1, t = . The sample size *n* will be 400 as 2500 is not possible.

Similarly for different value of *M* and *t*, *n* can be work out to be as

|  |  |  |  |
| --- | --- | --- | --- |
| *M*→  *t*↓ | 1 | 5 | 10 |
| ½ hr | 400 | 131 | 74 |
| 2 hrs | 156 | 40 | 21 |

**4.6 CLUSTER SAMPLING FOR PROPORTION.**

The idea of cluster sampling given in the previous sections, for estimation of population mean and total, can be easily extended for estimation of proportion. The notations for proportion in cluster sampling are given below:

 Proportion of specific character in ith cluster of population.

Proportion of the specific character in ith cluster of sample.

Using above notations, the unbiased estimator of population proportion may be defined as:

 (4.6.1)

The variance of *pc* may be written directly (using the concept of Chapter 2). From (4.4.1) when 

 (4.6.2)

An unbiased estimator of (4.6.2) can be readily written as:

 (4.6.3)

**4.7 CLUSTER SAMPLING WITH CLUSTERS   
OF UNEQUAL SIZES**

In the previous sections of this chapter we have discussed cluster sampling with cluster of equal sizes. In actual situations clusters i.e. block, households etc. ordinarily used in the design of sample survey do not contain the same number of elements. The estimation of population characteristics differ slightly in case of unequal cluster sizes. Let the population is divided into *N* cluster each having   
*M*1, *M*2, ......, *MN* elements. Then population mean is defined as:

 (4.7.1)

where 

When the clusters are of unequal size then for estimation of population total *Y* two methods can be used: (i) equal probability sampling (ii) probability proportional to size sampling.

In this chapter we shall consider only the sampling with equal probability. Probability proportional to size sampling will be discussed in Chapter 7. Various unbiased estimates can be formed but here we shall consider a few of them:

**4.7.1** , (4.7.2)

where  (4.7.3)

which is clearly unbiased.

The variance of  may be written directly by using Theorem (2.2).

 (4.7.4)

and an unbiased estimator of (4.7.4) is

 (4.7.5)

Equation (4.7.5) may be written as

 (4.7.6)

**4.7.2**  (4.7.7)

 (4.7.8)

The variance and unbiased variance are:

 (4.7.9)

and

 (4.7.10)

Since the variance depends on the variation of the product  and is therefore likely to be larger than . If *Mi* do not vary greatly then there is not much difference.

**4.7.3 Ratio-to-size estimate**

Let  (4.7.11)

Ratio-to-size estimate can be given as:

; (4.7.12)

where *Mi* are known. Here *Mi* are taken as bench mark variable.

We know that ,  may be written using the concept of Ratio estimate as:

 (4.7.13)

and approximate unbiased estimator for (5.7.13) is

 (4.7.14)

# EXAMPLE 4.5

A sample of 15 mohallahs was selected from a population of 200 mohallahs.   
The number of household and number of persons are recorded. Estimate   
total number of persons in 200 mohallahs. Find the standard error of this estimate.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mohallah | Number of  Households | Total Persons  yi | Average No.  of Persons |  |
| 1 | 25 | 250 | 10.00 | 100.00 |
| 2 | 30 | 260 | 8.67 | 75.169 |
| 3 | 35 | 218 | 6.23 | 38.813 |
| 4 | 39 | 300 | 8.11 | 65.772 |
| 5 | 22 | 216 | 9.82 | 96.432 |
| 6 | 30 | 288 | 9.60 | 92.160 |
| 7 | 24 | 230 | 9.58 | 91.776 |
| 8 | 70 | 540 | 7.71 | 59.444 |
| 9 | 67 | 520 | 7.76 | 60.218 |
| 10 | 55 | 537 | 9.76 | 95.258 |
| 11 | 70 | 515 | 7.36 | 54.170 |
| 12 | 38 | 210 | 5.53 | 30.581 |
| 13 | 47 | 270 | 5.74 | 32.948 |
| 14 | 67 | 410 | 6.12 | 37.454 |
| 15 | 53 | 401 | 7.57 | 57.305 |
|  |  |  | 119.536 | 987.500 |

# SOLUTION



The total number of persons in 200 clusters are



 may be calculated using expression (5.7.6) for this

 and = 952.814

.

= 0.3908



 = 78.18,

# EXAMPLE 4.6

In a forest nursery, there are six rows each of length 434 feet in the bed. To arrive at a suitable sampling unit for estimating the total number of seedlings in the   
bed, the entire population was studied using four types of sampling unit:   
(a) one-foot length of single row, (b) two-feet length of single row, (c) one foot of the complete width of the bed, and (d) two feet of the complete width of the   
bed. The results of this study are given. Find out the optimum sampling unit   
after comparing the relative cost-efficiencies of the four types of units considered here.

|  |  |  |  |
| --- | --- | --- | --- |
| Type of Unit | **Total No. of**  **Units N** | **Variance**  **Per Unit** | **Length of a row (in feet)**  **Covered in 15 Minutes** |
| **(1)** | **(2)** | **(3)** | **(4)** |
| One-Foot Row | 2604 | 2.537 | 44 |
| Two-Feet Row | 1302 | 5.746 | 62 |
| One-Foot Row | 434 | 23.094 | 78 |
| Two-Feet Row | 217 | 68.558 | 108 |

# SOLUTION

Since the entire population was studied using following types of sampling units:

1. one-foot length of single row
2. two-feet length of single row
3. one foot of the complete width of the bed and
4. two feet of the complete width of the bed.

There are six rows each of length 434 feet in the bed. Therefore, the relative size of the clusters will be

1, 2, 6 and 12 respectively.

Moreover the principal cost is of counting units, costs were estimated by time study. The relative values of *Ch* expressed as the time required to count one unit one.

 respectively.

All the results can be written in tabular form as:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type of Unit** | **Total**  **No. of**  **Units Nh** | **Variance**  **Per Unit** | **Length of a**  **Row**  **Covered in**  **15 Minutes** | **Relative**  **Size of Cluster**  ***Mh*** | **Cost in**  **15**  **Minutes**  ***Ch*** |
| One-Foot Row | 2604 | 2.537 | 44 | 1 | 1/44 |
| Two-Feet Row | 1302 | 6.746 | 62 | 2 | 2/62 |
| One-Foot Row | 434 | 23.094 | 78 | 6 | 6/78 |
| Two-Feet row | 217 | 68.558 | 108 | 12 | 12/108 |

We know that relative net prevision of a unit inversely proportional to the variance for fixed cost, i.e.

Relative net precision 

Therefore, relative next precision of the four units is

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 |
|  | 17.34 | 18.38 | 20.27 | 18.90 |

If we take 17.34 (minimum value) = 100 then relative precision may be expressed as

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 2 | 3 | 4 |
| 100 | 106 | 117 | 109 |

Since 3rd unit has maximum high precision so this is the optimum sampling   
unit.

# EXERCISE

1. A simple random sample of 15 clusters of 4 trees each, was selected   
   out of 308 bearing trees in a village of Sargodha District to study   
   the cultivation practices and yield of peaches. The yield in Kg. is given as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster/Tree | 1 | 2 | 3 | 4 |
| 1 | 5.53 | 4.84 | 0.69 | 15.79 |
| 2 | 26.11 | 10.93 | 10.08 | 11.18 |
| 3 | 11.08 | 0.65 | 4.21 | 7.56 |
| 4 | 12.66 | 32.52 | 16.92 | 37.02 |
| 5 | 45.98 | 5.07 | 1.17 | 6.53 |
| 6 | 0.87 | 3.56 | 4.86 | 27.54 |
| 7 | 54.21 | 34.63 | 52.55 | 37.20 |
| 8 | 1.24 | 35.97 | 29.54 | 25.28 |
| 9 | 37.94 | 47.07 | 19.64 | 28.11 |
| 10 | 54.92 | 17.69 | 26.24 | 6.77 |
| 11 | 7.13 | 34.35 | 12.18 | 9.86 |
| 12 | 25.52 | 38.10 | 24.74 | 1.90 |
| 13 | 14.23 | 16.89 | 25.90 | 20.70 |
| 14 | 3.53 | 40.76 | 5.50 | 1.25 |
| 15 | 45.98 | 5.17 | 1.17 | 6.53 |

1. Estimate the average yield per tree as well as the production   
   of orange in the village along with standard error of your estimate.
2. Estimate the intra-class correlation coefficient between trees within clusters.
3. Estimate the efficiency of cluster sampling as compared with simple random sampling.
4. From a population of 53 clusters a simple random sample of 14 clusters is selected; number of farms along with the number of cattle are given. Estimate the total number of cattle in the 53 clusters and find the variance.

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | No. of  Farms Mi |  | Average Cattle  Per Cluster |
| 1 | 19 | 66 | 3.47 |
| 2 | 28 | 326 | 11.64 |
| 3 | 28 | 392 | 14.00 |
| 4 | 29 | 350 | 12.07 |
| 5 | 31 | 331 | 10.68 |
| 6 | 41 | 351 | 8.56 |
| 7 | 46 | 697 | 15.15 |
| 8 | 51 | 586 | 11.49 |
| 9 | 53 | 739 | 13.94 |
| 10 | 55 | 914 | 16.64 |
| 11 | 61 | 619 | 10.15 |
| 12 | 64 | 784 | 12.25 |
| 13 | 83 | 906 | 10.92 |
| 14 | 80 | 1007 | 12.59 |

1. A sample of 20 mohallahs was taken from a population of 315 mohallahs and following information were recorded. Find the total number of   
   males and females in 315 mohallahs and find the standard error in each case.

|  |  |  |  |
| --- | --- | --- | --- |
| Mohallah | Households | Males | Females |
| 1 | 111 | 343 | 322 |
| 2 | 325 | 1036 | 921 |
| 3 | 207 | 764 | 411 |
| 4 | 846 | 4312 | 2412 |
| 5 | 189 | 557 | 209 |
| 6 | 409 | 1110 | 965 |
| 7 | 231 | 776 | 654 |
| 8 | 324 | 928 | 908 |
| 9 | 369 | 1204 | 1057 |
| 10 | 470 | 1929 | 1253 |
| 11 | 713 | 2469 | 2140 |
| 12 | 670 | 1424 | 1379 |
| 13 | 515 | 1391 | 1444 |
| 14 | 919 | 2854 | 1857 |
| 15 | 1479 | 5463 | 4458 |
| 16 | 837 | 2715 | 2359 |
| 17 | 1017 | 3689 | 3175 |
| 18 | 871 | 2670 | 2474 |
| 19 | 601 | 1874 | 1823 |
| 20 | 100 | 280 | 206 |

1. From a population of 100 clusters of 4 households each a sample of 10 clusters is taken. Find the total number of persons in the 100 clusters. And also find the standard error.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster | | Number of Persons | | | | |
| 1 | | 6 | | 5 | 2 | 6 |
| 2 | | 3 | | 8 | 7 | 5 |
| 3 | | 5 | | 6 | 3 | 8 |
| 4 | | 3 | | 8 | 4 | 5 |
| 5 | | 6 | | 7 | 3 | 3 |
| 6 | | 7 | | 8 | 3 | 4 |
| 7 | | 3 | | 3 | 5 | 4 |
| 8 | | 5 | | 6 | 3 | 4 |
| 9 | | 3 | | 5 | 6 | 7 |
| 10 | | 3 | | 7 | 8 | 2 |

1. Survey on pepper was conducted to estimate the number of pepper standards and production of pepper in a certain state. For this 3 cluster from 95 were selected by sample random sampling without replacement. The data are given as:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | Cluster Size |  | | | |
| 1 | 7 | 252, 386, 92, 293, 115, 59, 120 | |
| 2 | 11 | 41, 16, 19, 15, 114, 454, 212, 57, 28, 76, 199 | |
| 3 | 12 | 39, 70, 38, 37, 161, 38, 27, 219, 86, 128, 30, 20 | |

Estimate the total number of pepper standards along with standard   
error.

1. If the *NM* unit in a population are grouped at random to from *N* cluster of *M* units each. Show that the sample of *n* cluster with simple random sampling without replacement have the same efficiency as sampling of *nM* units with simple random sampling without replacement.
2. A population consists of *N* cluster each containing *M* elements. A simple random sample of *n* cluster is selected to estimate the population mean per element. An unbiased estimator would be  with variance  where  is defined in the text and finite population correction is ignored. Assuming , the within variance to be given by *aMg* where *g* > 0 and cost function to be , find the optimum size of cluster for which the variance of the estimator in minimum given the total cost of the survey, show that the optimum value of M will be smaller of *C*1 increases or *C*2 decreases.
3. Suppose in a study a cluster sample, of *n* clusters of *M* units   
   each was selected with srswr. Let *b* and *w* be unbiased estimates of between-cluster and within-cluster variances. Assuming the sample size in terms of the number of units to be fixed, obtain an estimate of the relative efficiency of cluster sampling as compared to that of direct sampling of units by estimating the sampling variances in the two cases unbiasedly.