***CHAPTER 3***

**STRATIFIED AND SYSTEMATIC RANDOM SAMPLING**

The previous chapter thoroughly explored one of the popular sampling designs known as simple random sampling. The simple random sampling provide strong basis for estimation of population characteristics when population units are relatively homogeneous with respect to characteristic under study. It is often be the case that the population units are heterogeneous with respect to characteristics under study or the population is naturally grouped. In both of these cases the simple random sampling does not provide proper basis for selection of the sample. Several other methods of sample selection are available that are useful in these sorts of situation. Two of the most popular methods namely Stratified and Systematic Random sampling are discussed in this chapter.

**3.1 THE STRATIFIED RANDOM SAMPLING**

When the population units are heterogeneous with respect to characteristic under study then it is useful to first divide the population into homogeneous groups. These homogeneous groups are called stratum. Stratification in a technique widely employed in survey sampling. It is the process of dividing the population into different strata, selecting a random sample from each stratum, and estimating the characteristics by using the selected sample. Finally estimate for each stratum are aggregated to produce an estimate for the whole population parameter. This division reduces the variability of the estimator thereby achieving the higher precision. There are number of reasons for using this type of sampling procedure. These reasons are illustrated below:

(i) it may increase precision by decreasing the variation within the stratum, (ii) information may be needed for individual stratum, (iii) it is easy to control the execution of survey, (iv) simultaneous work can be started by independent teams, and (v) sometimes in the selection of sample, different selection procedures may be used in different stratum.

The following are typical examples of stratified populations:

1. A population of school students stratified by school class or by sex.
2. A population of individual taxpayers stratified by province, by sex or by range of reported income.
3. A population of households in a country, each region being a separate stratum.
4. A population of retail establishments stratified by province by description (Grocer, butcher etc.) or by range of annual sales.

In stratification it is some time be the case that the stratification criteria are sharply defined (school class, sex, State), some stratification criteria admit some degree of subjective judgment (region, description of retail establishment) while a third group (age, range of reported income, range of annual sales) being quantitative in character, lead to arbitrary stratum boundaries.

Stratified random sampling has a long standing in the history of survey sampling. From the period 1926 to 1935 three significant were published on advances in Stratified Sampling. Bowley (1926) presented a classical method of sample selection from each stratum that was proportional to the total number of units in that stratum yielding a representative sample. Neyman (1934) showed how to allocate a sample among the strata in order to minimize the variance for a fixed total sample size. In the 1940’s the problem of partitioning a population into a strata was dealt especially with a principle of equi-partition and was used by the Indian Statistical Institute [Mahalanobis, 1952] and U.S. Bureau of Census [Hansen, Hurwitz and Madow, 1953]. Hagood and Barnet (1945) advocated the use of principal component analysis as a technique for determining the strata boundaries. Some choices of boundaries are better than others from the point of view of achieving an efficient sample design. The boundaries corresponding to the most efficient sample design for a particular item of interest lead to optimum stratification.

**3.2 NOTATION**

The following notation will be used in the context of Stratified random sampling.

If the *N* population units are divided into *k* strata then,

*N* = *N*1 + *N*2 + *N*3 + ---+ *Nh* + --- + *Nk*, where  is the number of population units in the hth stratum, similarly

*n* = *n*1 + *n*2 + *n*3 + ---+ *nh* + ----- + *nk*, where  in the number of sampling units in the hth stratum,

 = Population mean of the hth stratum,

= sample mean of the hth stratum,

= Population mean = 

 = Sample mean =  

, 

,  =  = Estimated population total.

The expectation and variance of the mean of stratified random sample are derived below:

**3.3 EXPECTATION OF SAMPLE MEAN**

The expectation is derived in the following theorem:

**THEOREM (3.1)**

In stratified random sampling without replacement the sample mean , is an unbiased estimator of population mean  i.e. , provided in each stratum  is unbiased for .

## PROOF

Taking expectation of the sample mean of stratified sample, we have:

 ◊ as .

Similarly it can be proved that 

The variance of the sample mean of stratified random sample is derived under various allocation methods in the following section:

**3.4. Variance and Unbiased Variance Estimator**

In this section we will derive the variance of  using following allocation methods:

1. Arbitrary Allocation,
2. Proportional Allocation, and
3. Optimum Allocation.

These are derived in the following theorems.

**3.4.1 Arbitrary Allocation**

In this type of allocation the total sample is allocated arbitrarilyamong the strata; strictly speaking there is no assumption for the allocation of sample units among the strata.

**THEOREM (3.2)**

The variance of sample mean, , for stratified random sampling for finite population sampling is

  (3.4.1)

**PROOF**

We know that the mean for stratified random sampling design is:

 

Applying variance expression, we have:

  (3.4.2)

Since within each stratum the simple random sampling design has been used, so we have:

  (3.4.3)

Using (3.4.3) in (3.4.2) we get:

  (3.4.1)

The expression (3.4.1) may be written as:

  (3.4.4)

Writing; the expression (3.4.4) can also be written as:

  (3.4.5)

For large value of *N*, above expression transformed to:

  (3.4.6)

The variance of  (estimated total) may be written in a straight-forward manner as:

  (3.4.7)

For large *N,* we have:

  (3.4.8)

Two special cases are of particular interest. The first, is Proportional allocation were *nh  Nh* and the other is optimum allocation, where *nh* are chosen either to minimize the variance of  given a fixed sample size n for fixed cost or cost is minimized for a given variance. We first consider the case of proportional allocation.

**3.4.2 Proportional Allocation**

If sampling fraction in all the strata is same, then the allocation is termed as proportional allocation. The sample size for hth stratum in this case is given as:

  (3.4.9)

This allocation was originally proposed by Bowley (1926) and is often used in practice because of its simplicity. This allocation is very useful if there is considerable difference between strata averages,  and not much difference in the strata variance. In case  are almost equal this allocation is not much useful as it brings a slight reduction in the variance of sample mean. In this allocation estimates become simple as strata weights are not required. This allocation is highly useful when stratification is done on the basis of geographic characteristics. This allocation is often (not invariably) appropriate for samples of persons. It is inefficient in circumstances where the population units differ greatly in size or importance, as is the case with surveys of retail establishment. In such circumstances, it is preferable to use optimum allocation.

**Theorem (3.3)**

If proportional allocation is used then the variance of mean of stratified random sample is:

  (3.4.10)

**PROOF**

Substituting the value of  for proportional allocation in (3.4.1) and on simplification

 

ignoring fpc, the expression (3.4.10) transformed to:

  (3.4.11)

The variance of estimated total  for proportional allocation may be obtained by multiplying the expressions (3.4.10) and (3.4.11) by N2.

**3.4.3 Optimum Allocation**

In the principle of optimum allocation *nh* are chosen either to minimize  for a fixed sample size *n* for fixed cost or cost is minimized for given variance. In general the aim of optimum allocation is to allocate *nh* in such a way that minimum variance is achieved for a minimum cost. The maximum precision may be achieved when the sampling units within each stratum are directly proportional to the stratum standard deviation. The two aspect of optimum allocation are

1. minimum variance for fixed cost
2. minimize cost for given variance

These aspects are discussed below:

1. **Minimum variance for fixed cost**

The sample size for hth stratum to minimize variance for a fixed cost in stratified random sampling is derived in the following theorem.

**Theorem (3.4)**

In stratified random sampling  will be minimum subject to the cost when *n*h is proportional to  i.e.

 .

**PROOF**

The variance of sample mean  for stratified random sampling is

  (3.4.3)

Let the simple cost function is:

 , (3.4.12)

where *C* = total cost, *C*0 = overhead cost, and *Ch* = cost per unit in the hth stratum.

The objective is to choose optimum value of *n*h, by minimizing, subject to given cost. For this we introduce Lagrange’s multiplier  i.e.



Partially differentiating w.r.t. *nh* and equating to zero, we have:

 

or  (3.4.13)

Summing (3.4.13) over all strata we have:

  (3.4.14)

Eliminating  from (3.4.13) and (3.4.14) we get

  (3.4.15)

If the cost is constant for all the strata i.e. *C*1 = *C*2 = … = *C* then the cost function will be

 *C* = *C*0 + *nC* (3.4.16)

The optimum sample size for the *h*th stratum from (3.4.12) in this case is:

  (3.4.17)

This is known as Neyman allocation after the name of Neyman (1934). Tchuprow (1923) gave the proof for optimum allocation which was discovered later.

The expression for minimum variance of  may be obtained by substituting the value of *nh* from (3.4.15) in (3.4.1).

  (3.4.18)

Ignoring fpc (3.4.18) becomes

  (3.4.19)

If the cost is constant then (3.4.18) transformed to:

  (3.4.20)

For large *N*, above expression becomes:

  (3.4.21)

Hence the variance of  is minimum when .

Stuart (1954) derived the same result by using Cauchy-Schwarz inequality as well.

**(ii) Minimum cost for the given variance**

If the variance is fixed, the choice of nh proportional to  must also minimize the total cost. Let .

Now using (3.4.1) and (3.4.16) the value of sample size (*n*) for fixed variance is

  (3.4.22)

This is the minimum sample size for estimating the mean with fixed variance. A special case of (3.4.22) may be obtained when the cost is the constant for each stratum. In this case (3.4.22) reduces to

  (3.4.23)

The main difficulty in using optimum allocation is that it requires the knowledge of *Sh*, which is difficult to obtain or sometime not available. In order to remove this difficulty *Sh* may be obtained from some previous survey. The use of this allocation becomes more difficult if more than one character is to be estimated from a single survey and in such cases this method may lead to loss of precision as compared to proportional allocation. However, if the characters are correlated, a gain in precision on the estimates of more important characters may still be secured by using this allocation. Neyman’s allocation concentrates the sampling efforts into those strata containing more variable population on units and thereby ensures minimum variance for given sample size or equivalently minimum sample size for given variance. The gain from using minimum variance allocation is especially large when sampling is done from highly skewed population. An example of such population is the population of retail stores, for which we want to estimate the mean sale . It would then be efficient to stratify the stores into say 3 strata i.e. small, medium and large, stores by some measure of size and select a small fraction *ns*/*Ns* of small stores, a large fraction *nM*/*NM* of medium stores and a still large fraction *nL*/*NL* of the large stores. Proportional allocation is identical to optimum allocation for fixed sample when each stratum is equally variable.

**3.4.4 Unbiased Variance Estimator**

Since simple random sampling procedure is applied within each stratum, hence for each stratum . Then an unbiased estimators of (3.4.1), (3.4.10), (3.4.18) and (3.4.20) may be obtained replacing  by  in the respective expressions.

**3.5 STANDARD ERROR AND CONFIDENCE LIMITS**

The standard error may be obtained by taking the square root of the variance of  and the confidence limits for mean and estimated totals are

  (3.5.1)

and  (3.5.2)

**3.6 relative precision of simple random
sampling and stratified random sampling with different allocations**

In general the stratification brings a reduction in the variance of sample mean than simple random sampling. In this section comparisons have been made between different allocations and with simple random sampling.

**3.6.1 Simple Random Sampling and Proportional Allocation**

From (2.4.3) and (3.4.10)

 

 

Adding and subtracting  in the first term of the right hand side and simplifying, we get



 =  (3.6.1)

 = 

which need not necessarily be positive, it follows that

 

provided

  (3.6.2)

If *Nh* is large *Nh* – 1 ~ *Nh* and 1- *n*/*N* is taken as unity (ignoring f.p.c), then (3.6.1) is

  (3.6.3)

or  (3.6.4)

This difference is non negative, being proportional to the weighted variance of the  with weight *Wh*. Consequently, given this type of sampling and estimation, it is impossible to loose efficiently as a result of stratification and the greatest efficiency is achieved when the stratum means differ from each other as much as possible.

The difference in (3.6.4) is non negative, giving  Equality holds only when. Further

 , (3.6.5)

which shows that the ratio of proportional allocation to random sampling does not depend on the size of sample.

**3.6.2 Proportional and Optimum Allocations**

Consider next the case of optimum allocation where the cost of sampling does not differ from stratum to stratum, and proportional allocation if term 1/*Nh* in all the strata is negligible then, from (3.4.1) and (3.4.2)

  (3.6.6)

 The expression in the braces is the weighted variance of the unit weights *Wh*. Thus this form of optimum allocation has two reductions in variance compared with simple random sampling, one term being proportional to the weighted variance of the stratum mean and the other proportional to the weighted variance of the stratum standard deviation. The weights in each case are the proportions of population units in the strata. We can write (3.6.6) as:-

 

 

  (3.6.7)

Quantity in (3.6.7) is always positive so 

**3.6.3 Optimum Allocation and Simple Random Sampling**

From (3.6.3) and (3.6.7) we have

 

 (3.6.7)

Since the right hand side is the sum of two non-negative terms, hence

 

We conclude if f.p.c. is ignored then

  (3.6.9)

**3.6.4 Simple Random Sampling and Arbitrary Allocation**

The difference in variance of  and  under arbitrary allocation is:

  (3.6.10)

which is also positive, hence if  is taken proportional to *Wh* as the first term of (3.6.10) reduces to zero. Also if *nh* is proportional to  then .

**3.6.5 Loss in Precision due to Failure in Achieving an Optimum Allocation**

Given *n*, the sample size for optimum allocation is

  (3.6.12)

The variance of sample mean under optimum allocation is

  (3.4.20)

Suppose *nh* is the sample size used in *h*th stratum for arbitrary allocation then

  (3.4.3)

Increase in variance caused by imperfect allocation is:

  (3.6.13)

From (3.6.12), we have:

 (3.6.14)

Using (3.6.14) in the first term of (3.6.13), we get



Since , we then have



  (3.6.15)

Now  =

Hence  (3.6.16)

Ignoring f.p.c. the relative increase in variance is

  (3.6.17)

If f.p.c. is not ignored then relative increase in variance is

  (3.6.18)

It is not easy to realize the fractional implication of this result by just looking at the formula. Some numerical results are required

 Let g = greater value of  found in any of the stratum

Then the greater relative deviation from the optimum size is

 

If g = 0.02 then the proportional increase cannot exceed more than (.02)2 = 4%. This rough rule usually over estimate the true increase in relative variance because we substitute g for any relative deviation of the optimum sample size. To illustrate suppose we have 7 strata with  given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stratum |  |  |  |  |
| 1 | 30 | 19 | 0.367 | 4.033 |
| 2 | 46 | 36 | 0.217 | 2.174 |
| 3 | 38 | 34 | 0.105 | 0.421 |
| 4 | 31 | 39 | 0.258 | 2.064 |
| 5 | 17 | 24 | 0.412 | 2.882 |
| 6 | 13 | 17 | 0.308 | 1.231 |
| 7 | 25 | 31 | 0.248 | 1.44 |
|  | 200 | 200 |  | 14.615 |

Since  is 0.412 in stratum 5, so the rough rule gives us that
the proportional increase is about (.412)2 = 17% whereas actual increase is
14.615/200 = 7.3%.

**EXAMPLE 3.1**

Population of 468 villages of Multan district was divided into 11 strata. The means, variances and standard deviations of each stratum are given below:

|  |  |  |  |
| --- | --- | --- | --- |
| Range |  |  | *Sh* |
| 1 – 499 | 261.59 | 21960.25 | 148.19 |
| 500 – 972  | 707.18 | 15996.98 | 126.48 |
| 973 – 1617  | 1312.97 | 31163.40 | 176.53 |
| 1618 – 2310 | 1565.74 | 37217.60 | 192.92 |
| 2311 – 2899 | 2598.32 | 36297.48 | 190.52 |
| 3000 – 3680 | 3305.4 | 62846.85 | 250.69 |
| 3681 – 4278 | 3962.07 | 32945.03 | 181.51 |
| 4279 – 4959 | 4682.83 | 34680.49 | 186.23 |
| 4960 – 6094 | 5498.53 | 143691.2 | 379.07 |
| 6095 – 6769 | 6773.41 | 40006.63 | 200.02 |
| 6770 – 7951 | 7289.17 | 138856.5 | 372.63 |

A sample of size 46 is drawn from 468. On the basis of above information Estimate the sample size for proportional and optimum allocations.  and .

**SOLUTION**

For proportional allocation  whereas for optimum allocation . The necessary calculations are given below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Wh* | *nh*(Prop)=*nWh* | WhSh |  |  |  |
| .15 | 7 | 22.23 | 5 | 3294.04 | 729078.9 |
| .15 | 7 | 18.97 | 4 | 2399.55 | 464149.1 |
| .16 | 7 | 28.24 | 7 | 4986.14 | 212808.8 |
| .13 | 6 | 25.08 | 6 | 4838.29 | 32566.3 |
| .09 | 4 | 17.15 | 4 | 3266.77 | 1569.8 |
| .06 | 3 | 15.04 | 4 | 3770.81 | 42214.1 |
| .06 | 3 | 10.89 | 3 | 1976.70 | 134248.6 |
| .05 | 2 | 9.31 | 2 | 1734.02 | 245661.3 |
| .06 | 3 | 22.74 | 5 | 8621.47 | 551683.3 |
| .05 | 2 | 10.00 | 2 | 2000.33 | 208866.6 |
| .04 | 2 | 14.91 | 4 | 5554.26 | 930422.3 |
|  |  |  |  | 42442.39 | 4147269.2 |

Now 

 

If fpc is ignored then:

 , and 

The variance and standard error under proportional allocation is:

;

If fpc is ignored then:

 ; 

The variance and standard error under optimum allocation are:

 

 ; .

If fpc is ignored

 ; 

If fpc is ignored

 

 = 922.66 + 90158.03 = 91080.69

**Example 3.2**

The smoking information given in the following table was obtained using a census questionnaire of the adult male population in a Australian City during 1966.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Type of smoking | stratum | adult male population size | Daily average No. of cigarette smoked | Standarddivision |
| Daily rate in cigarette | h | Nh |  |  |
| Light smoker < 10 | 1 | 28,900 | 8.1 | 8.63 |
| Medium smoker between (10-20) | 2 | 38,300 | 15.3 | 16.30 |
| Heavy smoker > 20 | 3 | 52,800 | 28.2 | 30.04 |
| Total |  | 120,000 |  | 23.88 |

In order to examine the current smoking habits of adult males in the city, a sample survey of size 800 is planned for 1968 using the information obtained in the 1966 census result.

Obtain the variance of the estimates of the total and average number of cigarettes smoked per day by using:

1. Simple random sampling.
2. Stratified modern sampling with proportional allocation.
3. Stratified random sampling with optimum allocation.

Check the difference between the variances using the standard formulae.

(Source New South Wales University).

**SOLUTION**

**i) Simple Random Sampling**

 = = = 0.707698

 **ii) Proportional Allocation**

 We have for proportional allocation: *nh*= 

 Therefore *n*1­ = 193, *n*2 = 255 and *n*3 = 353

 so  = 

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *h* | *Nh* | *Wh* |  | *Sh* | *NhSh* | *WhSh* |  |
| 1 | 28900 | 0.2408 | 74.4769 | 8.63 | 249407.00 | 2.078 | 17.934 |
| 2 | 38300 | 0.3192 | 265.6900 | 16.39 | 624290.00 | 5.203 | 85.747 |
| 3 | 52800 | 0.4400 | 902.4016 | 30.04 | 1586112.00 | 13.218 | 397.657 |
|  | 120000 |  |  |  | 2459809.00 | 20.499 | 501.338 |

  = 0.62058

**iii) Optimum Allocation**

 Under optimum allocation we have: *nh* = 

 Therefore *n*1 = 193, *n*2 = 255 and *n*3 = 352

  = 0.521081

**Example 3.3**

If the cost function is  then find the optimum allocation of *nh* to minimize  for fixed cost (Cochran).

**SOLUTION**

Using the Lagrangian multiplier, the function to be optimize is:

 

Partially differentiating w.r.t. *nh* and equating to zero we have:

  (A)

Summing over *n*h, we have:

  (B)

Comparing (A) and (B), we have:

 

Hence *n*h is proportional to 

**3.7 STRATIFIED SAMPLING FOR PROPORTION.**

Stratified Random Sampling can effectively be used for estimation of population proportion.

An unbiased estimator for population proportion in stratified random sampling is

  (3.7.1)

In order to find the variances etc. for proportion in stratified random sampling, the results of Chapter 2 may be used. For single stratum they are

  (3.7.2)

**3.7.1 Variance and Unbiased Variance Estimator for Arbitrary Allocation**

We can easily apply the results of (3.7.2) to Section (3.4) to find the variance and an unbiased variance estimator of proportion in stratified random sampling

The variance of *pst* for arbitrary allocation is

 

Analogically to Chapter 2, we have:

  (2.7.1)

Hence  (3.7.3)

If *Nh* – 1 ~ *Nh* and *nh*/*Nh* is ignored in (3.7.3) then

  (3.7.4)

The variance estimators of (3.7.3) and (3.7.4) are

  (3.7.5)

And  (3.7.6)

**3.7.2 Variance of Proportional and Optimum Allocations**

A population is divided into *k* strata each stratum have *N*1, *N*2, …. *Nk* population units. If f.p.c. is ignored then the variances of stratified random sampling for proportional and optimum allocations for overall proportion of population members possessing the particular characteristics are approximately

 

and .

Since 

Substituting this value is (3.4.10) we obtain

 

For large *N*; above expression becomes:

  (3.7.5)

Similarly putting the value of  in (3.4.18) we get

 

  (3.7.6)

For large *N*, second term of the above expression is approximately zero. So the variance of proportion of stratified random sample is:

  (3.7.7)

If the cost is same for each unit then (3.7.7) is

  (3.7.8)

Variance estimators may be obtained by replacing *PhQh* with *phqh* in the respective expressions.

**COROLLARY**

Following expression may be proved in a straightforward manner by putting  in the respective expressions for variance comparison:

  (3.7.9)

and (3.7.10)

**Example 3.6**

Use the following information and compare *Var*(*pst*) when f.p.c. is ignored for proportional and optimum allocations for fixed sample size *n* when each stratum is of equal size.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stratum | 1 | 2 | 3 | 4 |
| *Ph* | 0.1 | 0.3 | 0.6 | 0.8 |

**solution**

Since 

And 

The relevant calculations are given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stratum | *Ph* | *Qh* | *PhQh* |  |
| 1 | 0.1 | 0.9 | 0.09 | 0.300 |
| 2 | 0.3 | 0.7 | 0.21 | 0.458 |
| 3 | 0.4 | 0.4 | 0.24 | 0.490 |
| 4 | 0.8 | 0.2 | 0.16 | 0.400 |
|  | 0.70 | 1.648 |

Since each stratum is of equal size, so we have:

 

 

The reduction in variance due to optimum allocation is

 

**Example 3.7**

In a city area 2300 households was divided into 5 strata on the basis of their monthly income. Simple random sampling technique was applied to select a sample within each stratum and information was obtained about the renting house. Find the proportion of households living in rented houses; also find total number of houses in the population on rent and calculate.

|  |  |  |  |
| --- | --- | --- | --- |
| Stratum(income) | *Nh*Number of households | *nh*Number of Households in sample | Renting houses |
| < 50 | 1190 | 50 | 30 |
| 50 – 100 | 520 | 38 | 18 |
| 100 – 200 | 350 | 30 | 10 |
| 200 – 400 | 180 | 40 | 8 |
| 400 and over | 60 | 20 | 4 |
| Total | 2300 | 178 | 70 |

**SOLUTION**

Proportion of the houses in each stratum in the sample is

|  |  |  |
| --- | --- | --- |
| Stratum | *Ph* | *Qh* |
|  | 0.60 | 0.40 |
| 2 | 0.47 | 0.53 |
| 3 | 0.33 | 0.67 |
| 4 | 0.20 | 0.80 |
| 5 | 0.20 | 0.80 |

Proportion of the rented household in the population is

 

Total number of household on rent in the population is *Npst* = 1122.

The variance of *pst* is:

  

so  = 9522

and  = 97.5807

**3.8 ESTIMATION OF GAIN IN PRECISION DUE TO STRATIFICATION**

Sometimes in a survey it is useful or of interest to find out how stratification is effective as compared to simple random sampling. In comparing the precision of stratified sampling with simple random sampling it is assumed that the population values of mean  and variances  are known. In order to estimate the gain in precision due to stratification, an estimate of the variance of the estimates in case of simple random sampling is obtained from a sample and a comparison can be made with a situation in which no stratification is done. The main problem is to get an unbiased estimate of  based on given stratified sampling. The variance estimator of  given by J.N.K. Rao (1962) is:

  (3.8.1)

It can be easily shown that

  (3.8.2)

and 

or  (3.8.3)

The efficiency of the relative gain in precision due to stratification is obtained as

  (3.8.4)

**Example 3.8**

The following data are derived from a stratified random sample of tyre dealers. The dealers were assigned to strata according to the number of new tyres held at the previous census. The sample mean  are the mean number of new tyres per dealer. Estimate the gain in precision due to stratification. (Source Cochran 1977).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stratum |  *Nh* |  *Wh* |   |   |   |
| 1. | 19850 | 0.8032 | 4.1 | 34.8 | 3000 |
| 2 | 3250 | 0.1315 | 13.0 | 92.2 | 600 |
| 3 | 1007 | 0.0407 | 25.0 | 174.2 | 340 |
| 4 | 606 | 0.0245 | 38.2 | 320.4 | 230 |
| Total | 24713 |  |  |  | 4170 |

**Solution**

The relative calculations are:

|  |  |  |  |
| --- | --- | --- | --- |
| \* |  |  |  |
| 154795.2 | 1024228.20 | 0.0074835 | 27.9514 |
| 156627.8 | 848400.58 | 0.0026572 | 12.1243 |
| 271553.8 | 804278.46 | 0.0008487 | 7.0899 |
| 408996.8 | 1077617.70 | 0.0008361 | 7.8498 |
| Total: | 3754524.94 | 0.0118256 | 55.0154 |

  = 

  = 

  =  (ignoring f.p.c)

and  = 

So the estimated gain in precision due to stratification (using 3.8.4) is

 = 110%

 = 0.0131931

so the estimated gain due to stratification with proportional allocation is

  = 0.88225 ~ 88.2%

**EXAMPLE 3.9**

A finite population of size *N* with parameters  has two strata of size *N*1 and *N*2 where the first stratum has all zero values, so  then

 

**SOLUTION**

 

We know that

 

since  therefore

 

 

since *N* = *N*1 + *N*2 therefore

 

**3.9 SYSTEMATIC SAMPLING**

Systematic sampling is operationally more convenient than simple random sampling. This selection procedure is different from simple random sampling procedure in the context that in simple random sampling procedure, every unit is selected with the help of random numbers whereas in systematic sampling only the first unit is to be selected at random and the remaining units are automatically determined by the skip interval. The systematic sampling can be illustrated as under:

Suppose there are *N* units in the population numbered from 1 to *N*. If *N*/*n* = *k*, where n denotes the number of units in the sample and *k* (skip interval) is an integer, the population of *N* units is divided into *n* groups each containing *k* units. If *r*th (say) unit is selected at random from the first group of *k* units, then (*k*+*r*)th unit, (2*k*+*r*)th units are selected from the second and third group respectively and so on till the sample size of *n* units is selected. The random number chosen from the first group of *k* units is known as **random start** and *k* is termed as **skip interval**.

If there are N units in the population and N/n = [k] (an integer greater than k) then the systematic selection is explained as:

### Table 3.1

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Sample Composition | Probability*P*(*s*) | SampleMean |
| 1 | 1, *k*+1, 2*k*+1, …, (i-1)*k*+1, … (*n*-1)*k*+1 | 1/*k* |  |
| 2 | 2, *k*+2, 2*k*+2, …, (*i*-1)*k*+2, … (*n*-1)*k*+2 | 1/*k* |  |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| *r* | *r*, *k*+*r*, 2*k*+*r*, (*i*-1)*k*+*r*, … (*n*-1)*k*+*r* | 1/*k* |  |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| *k* | *k*, 2*k*, 3*k* ,… ,*ik* …. *Nk* | 1/*k* |  |

The probability of selecting the group is 1/*k* which is in fact the probability with which any member of the group is selected in the sample. It is commonly applied when population frame is not available or not possible.

Systematic selection is useful in survey of forest trees; in horticultural experiments it becomes most important. It is also very useful if some one is interested to measure satisfaction of patients in an outdoor ward of any hospital or to measure satisfaction of clients of any bank, as in these cases frame is not available. This procedure was employed in Indian population census during 1941, and was also used in households surveys in U.S.A. in 1960. Use of systematic sampling was considered by Hajek (1942), Finny (1948), Cochran (1946) and Yates (1948). Perhaps first time Madow and Madow (1944) developed the mathematical theory of this selection procedure.

It is usual to analyze results from systematic sample as though they are from a simple random sample of the same size. If the population from which the sample is selected is effectively in random order to begin with, this is a fair enough assumption. But if there is some structure running through the population list, it can be quite dangerous to use this assumption.

Suppose, first that there is some variable (say income) which intends to be high at the start of the list and low at the end. Then the systematic sampling procedure becomes artificial stratification. Every group of *k* units is a pseudo-stratum, from which one unit is selected. This virtual stratification ensures that the variance
of the sample estimator is smaller, sometime much smaller, than that from a simple random sample of the same size. However, the value of *s*2 calculated from

  (2.5.6)

will be much of the same size as – in facts usually slightly larger than – the corresponding value from a random sample. The variance is therefore, over estimated. One way of reducing this over estimation is to modify the formula (Brewer 1969) of *s*2 so as to exclude from considerations all between units other than contiguous sample units i.e.

 (3.9.1)

The factor *n* disappears from the denominator because there are only (*n* – 1) comparisons instead of the *n*(*n* – 1) there previously.

The opposite effect occurs when there is a structure in the population list which has the same period as the skip interval *k*. As an example, a list of soldiers in which every 10th is a sergeant, subjected to a skip interval 20, will result in a sample which contains all sergeants or none. The true variance will be far higher than the simple random sampling variance, because the sample units are similar to each other. An estimate of the variance will be small. In this case, it would not help to use modified expression (3.9.1). There is no way of remedying the situation.

In summary, the following remarks are useful for systematic sampling procedure:

1. Selection is simple, quicker and easier.
2. It involves less cost as compared to simple random sampling.
3. A complete and up to date frame is not strictly needed but the idea of the population is necessary.
4. Time spent on actual selection of sample is much less than simple random procedure.
5. It often gives the advantages of stratification.
6. The variance estimated is some what higher than of a simple random sample of the same size.
7. The estimate of variance is different if the arrangement of population units is changed.

**3.10 EXPECTATION OF SAMPLE MEAN.**

In this section we have derived the expectation of the mean of a systematic sample by considering two situations. The first situation that we have considered is that when *N*=*nk*. The expectation is derived in the following theorem.

**THEOREM (3.5)**

In a systematic sample of size *n* drawn from a finite population of *N* units, when *N = nk*, where *k* is an integer, the sample mean  is an unbiased estimator of population mean, i.e. 

From table 3.1 it is clear that:

 , (3.10.1)

and , (3.10.2)

then . (3.10.3)

# Proof

# If we consider all the *k* sample then,

  (3.10.4)

as the probability of selection of the *r*th systematic sample is 1/*k*.

We know that

  (3.10.2)

Substituting  in (3.10.4), we get

 

Hence  is an unbiased estimator of. ◊

When *N*/*n* is not an integer, the sample does not remain fixed as it depends on the random start. In these situations slight modifications are suggested to obtain unbiased estimates. These modifications are given below:

1. Suppose we have a population of 7 units, say *Y*1, *Y*­2, …, *Y*7, with *k* = 3. If the random start is the first number the sample units will be *Y*1, *Y*4, and *Y*7, but if the random starts are 2nd and 3rd units, the sample will be either *Y*2, *Y*5 or *Y*3, *Y*6 respectively with probability 1/3 in each case. An unbiased estimator may be obtained as

 

 In this situation since *k* = 3 and *N* = 7

 

1. Select a random start from 1 to *N* [instead of 1 to *k*] and take every *k*th unit both in backward and forward direction. If in the above population the random start from 1 to 7 is 4 then the sample is *Y*1, *Y*4, and *Y*7 with probability 3/7, if the random start is second or third unit the samples will be *Y*2, *Y*5 and *Y*3, *Y*6 respectively with probability 2/7 in each case. An unbiased estimator of population mean is

 

This procedure was suggested by Cochran (1963). Estimated population total in systematic sampling is simply obtained by multiplying mean with *N*.

**3.11 VARIANCE AND VARIANCE ESTIMATOR
OF SAMPLE MEAN**

In this section we have derived the variance of systematic sample under the condition that sample size is a multiple of population size. The variance is derived in the following theorem:

**THEOREM (3.6)**

In systematic sample of size *n*, selected from a population of *N* units, the variance of is

  (3.11.1)

or  (3.11.2)

where  (3.11.3)

**Proof**

Since there are *k* possible samples,  is the mean of *r*th systematic sample, therefore,

 

To prove (3.11.2) let us define *S*2 as,

 

Adding and subtracting  in the right hand side of above expression and on simplification, we have:

  (3.11.4)

Using (3.11.3), in (3.11.4) we have

  (3.15.5)

Using (3.15.1) in (3.15.5), we get

 ◊

The total variance *S*2 may be divided in terms of variance between various systematic sample means, , and variance within a specific systematic sample,  From (3.11.4)

  (3.11.6)

From analysis of variance technique we can write (3.11.6) as

  (3.11.7)

Note that, the variance of the estimated total,  may be written in a straight forward way:

 

or  (3.11.8)

The two components of total variance play dominant role in deciding about the precision of a systematic sample as compared with the simple random sample. This is proved in the following theorem.

**THEOREM (3.7)**

The systematic sample is more efficient than simple random sample if the variation within the systematic sample is more than the total variation.

**Proof:**

We know that

  (2.4.1)

and  (3.11.2)

Comparing (2.4.1) and (3.11.2) we get

 

if  (3.15.9)

The variance of sample mean of systematic sample may also be expressed in terms of intra-class correlation coefficient, ρ as under:

We know that 

Substituting the value of  in the above expression we get

 

 

Or  (3.11.10)

where ρ is the correlation coefficient between pairs of units that are in the same systematic sample i.e.

  (3.11.11)

Since *S*2 is fixed, the value of ρ should be negative to reduce the variance. This is possible when the arrangement of units within each systematic sample are heterogeneous. So the variance of systematic samples does not only depends on *S*2 and *n* like simple random sampling but also depends on ρ which generally varies with the sample size and the arrangement of units. Since  is never less than zero, ρ cannot be less than –1(n-1), hence ρ must lies between –1/(n-1) and 1.

The relative precision of systematic sample and simple random sample in term of ρ is obtained below by using (3.11.10) and (2.4.1):

  (3.11.12)

It is clear that,

 

 

and 

If ρ = 1 then = (*N*-1)/ (*k*-1).

Further suppose  is the mean of the *i*th stratum i.e.

  (3.11.13)

The stratum mean square within *i*th stratum is

  (3.11.14)

The pooled mean square between units within stratum is

  (3.11.15)

as each of the *n* strata contribute *K*–1 degree of freedom. Then,

  (3.11.16)

Using (3.11.14)

  (3.11.17)

The variance of  may be written as:

 

 

 

 

 , (3.11.18)

Where 

Comparing (3.11.17) and (3.11.18), we have

  (3.11.19)

We can immediately see that the relative efficiency of systematic sampling over stratified random sampling depends upon the value of ρw­.

 

 

 

**3.11.1 Variance Estimator**

Since the systematic sampling procedure does not ensure the inclusion of each  pairs of units of the population at least in one of the samples, which is a necessary condition for an unbiased estimator, hence under systematic sampling procedure it is not possible to obtain an unbiased variance estimator from one sample. If, however, two or more systematic samples are drawn with different random start then combined variance estimator is possible. Following variance estimate of sample mean,  and estimated total  may be suggested:

  (3.11.20)

and  (3.11.21)

As we know that

 

Now there are *n*(*n*-1) terms in the summation  each of which has the same expectation, i.e. (yi – yj)2 has the expectation *S*2. When we pass from simple random sampling without replacement to systematic sampling only the expression  should be taken into account, and even they would over estimate the variability within a quasi-stratum (being th part of the population).

**EXAMPLE 3.10**

Following data relating to the height of some plants in a forest with *N* = 24,
= 75, 72, 62, 52, 52, 81, 50, 56, 57, 57, 71, 81, 57, 43, 70, 49, 44, 57, 48, 54, 49, 36, 37, 45, draw all possible samples as:

1. *k* = 3 and *n* = 8
2. *k* = 4 and *n* = 6
3. *k* = 5,

Find the mean and variance of  in each case also prove that sample mean is an unbiased estimator of population mean.

 

**SOLUTION**

Case (i): *k* = 3 and *n* = 8 following are the possible samples.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Observations | Total | Mean |
| 1 | 75 | 52 | 50 | 57 | 57 | 49 | 48 | 36 | 424 | 53.0 |
| 2 | 72 | 52 | 56 | 71 | 43 | 44 | 54 | 37 | 429 | 53.63 |
| 3 | 62 | 81 | 57 | 81 | 70 | 57 | 49 | 45 | 502 | 62.75 |
|  |  |  |  |  |  |  |  |  | 1355 | 56.46 |

 

 

where as

 

Case (ii): k = 4 and n = 6

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Observations | Total | Mean |
| 1 | 75 | 52 | 57 | 57 | 44 | 49 | 334 | 55.67 |
| 2 | 72 | 81 | 57 | 43 | 57 | 36 | 346 | 57.67 |
| 3 | 62 | 50 | 71 | 70 | 48 | 37 | 338 | 56.33 |
| 4. | 52 | 56 | 81 | 49 | 54 | 45 | 337 | 56.17 |
|  |  |  |  |  |  |  | 1355 |  |



Case (iii): *k* = 5 and *n*1 = *n*2 = *n*3 = *n*4 = 5 and *n*5 = 4

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Observations | Total | Mean |
| 1 | 75 | 81 | 71 | 49 | 49 | 325 | 65.0 |
| 2 | 72 | 50 | 81 | 44 | 36 | 283 | 56.6 |
| 3 | 62 | 56 | 57 | 57 | 37 | 269 | 53.8 |
| 4 | 52 | 57 | 43 | 48 | 45 | 245 | 49.0 |
| 5 | 52 | 57 | 70 | 54 | - | 233 | 58.25 |
|  |  |  |  |  |  | 1355 |  |

. The variance of  may also be computed like cases (i) and (ii). If we find  like case
(i) and (ii) it will be  = (65.0 + 56.6 + 53.8 + 49.0 + 58.25)/5 = 56.53 which is not equal to the population mean 56.46. The variance of systematic and simple random sampling may also be obtained using analysis of variance technique.

|  |
| --- |
| **Case (i)** |
| Source of Variance | d.f | S.S | M.S.S |
| Between samples | 2 | 476.59 | 238.29 = S2b |
| Within samples | 21 | 3275.17 | 155.96 = S2w |
| Total  | 23 | 3751.76 | 163.12 = S2 |

 

 

 

|  |
| --- |
| **Case (ii)** |
| Source of Variance | d.f | S.S | M.S.S |
| Between samples | 3 | 13.13 | 4.38 = S2b |
| Within samples | 20 | 3738.83 | 186.94 = S2w |
| Total  | 23 | 3751.76 | 163.12 = S2 |

 

In case (i)  and in case (ii) as  and this satisfies the Theorem (4.3).

Likewise the comparison of stratified sampling and systematic sampling depends on the properties of population. It is difficult to make a general rule for comparison. If the order of the population units is changed, different groups will be formed; as a result the variance will be changed i.e. if a population has 6 units 1, 2, 3, 4, 5, and 6 with n = 2 the  = 1.667 and different systematic samples will be (1, 4), (2,5) and (3, 6) the variance of systematic sampling is  = 0.667. If the arrangement is changed as 1, 4, 5, 2, 3, and 6 the possible systematic samples are (1, 2), (4, 3) and (5, 6). In this case the variance of systematic sampling becomes 2.667.

**EXAMPLE 3.11**

Following is the population of 70 villages along with cultivated area (in acres). Select a sample of size 10 with systematic sampling method and estimate the total population and cultivated area of the villages. Find the standard error for the estimate.

**SOLUTION**

We have *N* = 70,  and *n* = 10

Since *k* = 7, we select a sample with a random start of 6, the villages in the sample are selected as 6 + 7, 6 + 2 x 7, …… The sample consists of the villages shown in Table 3.2. The population and cultivated area are given against each village in the table below:

**Table 3.2**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S# | No. Villages | Population | CultivatedArea | Population(*y*i – *y*i+1)2 | Cultivated Area (*y*i – *y*i+1)2 |
|  | 6 | 2282 | 1110 | -- | -- |
|  | 13 | 5201 | 1840 | 8520561 | 532900 |
|  | 20 | 1607 | 1225 | 12916836 | 378225 |
|  | 27 | 1567 | 970 | 1600 | 65025 |
|  | 34 | 773 | 602 | 630436 | 135424 |
|  | 41 | 828 | 277 | 3025 | 105625 |
|  | 48 | 547 | 372 | 78961 | 9025 |
|  | 55 | 726 | 636 | 32041 | 69696 |
|  | 62 | 1225 | 634 | 249001 | 00004 |
|  | 69 | 663 | 422 | 315844 | 44944 |
|  |  | 15419 | 8088 | 22748305 | 1340868 |

|  |
| --- |
| Estimated mean and total |
|  | Population | Cultivated Area |
|  | 15419 | 8088 |
|  | 15419 | 808.8 |
|  | 107933 | 56616 |

Variance estimator and standard error of mean and total are:

|  |  |  |
| --- | --- | --- |
|  | Population | Cultivated Area |
|  | 22748305= 1263794.722 | 1340869=74492.667 |
|  | 108325.262 | 6385.086 |
|  | 530793783.8 | 31286921.4 |
|  | 329.128 | 79.907 |
|  | 23038.962 | 5593.471 |

**EXAMPLE 3.13**

From a population given in Appendix 2 relating to population of 523 villages, four systematic samples of size 43 have been drawn and the data regarding these samples have been given. Calculate the average population of each village along with their standard errors and compare standard errors of these samples.

Population mean 

 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  | SE |
| Sample I | 3263.7 | 239579.630 | 489.470 | 501.27 |
| Sample II | 3440.5 | 341938.717 | 584.750 | 517.44 |
| Sample III | 3731.8 | 430879.625 | 656.414 | 652.44 |
| Sample IV | 4286.5 | 355611.591 | 596.330 | 584.81 |

**Table 3.4**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sample 1 | Sample 2 | Sample 3 | Sample 4 |
|  | *yi* |  | *yi* |  | *yi* |  | *yi* |  |
|  | 701 | 961 | 1071 | 128164 | 428 | 94249 | 793 | 14500864 |
|  | 670 | 5625 | 713 | 54184321 | 121 | 779689 | 4601 | 3345241 |
|  | 745 | 30976 | 8074 | 54449641 | 1004 | 24098281 | 2772 | 5076009 |
|  | 921 | 3912484 | 695 | 14417209 | 5913 | 5184 | 5025 | 2059225 |
|  | 2899 | 2795584 | 4492 | 11675889 | 5841 | 4468996 | 6460 | 19838116 |
|  | 1227 | 874225 | 1075 | 77193796 | 7955 | 24641296 | 2006 | 4739329 |
|  | 292 | 4334724 | 9861 | 69172489 | 2991 | 6906384 | 4183 | 4853209 |
|  | 2374 | 57395776 | 1544 | 902500 | 363 | 32844361 | 6386 | 9114361 |
|  | 9950 | 55726225 | 594 | 163737616 | 6094 | 324108009 | 9405 | 68807025 |
|  | 2485 | 63043600 | 13390 | 179024400 | 24097 | 553566784 | 1110 | 18740241 |
|  | 10425 | 53743561 | 10 | 20675209 | 569 | 58936329 | 5439 | 2022084 |
|  | 3094 | 60372900 | 4557 | 17935225 | 8246 | 49730704 | 4017 | 160782400 |
|  | 10864 | 32970564 | 322 | 1238769 | 1194 | 15124321 | 16697 | 92833225 |
|  | 5122 | 19686969 | 1435 | 1267876 | 5083 | 5822569 | 7062 | 34656769 |
|  | 685 | 3069504 | 309 | 537289 | 2670 | 5331481 | 1175 | 1763584 |
|  | 2437 | 5579044 | 1042 | 15523600 | 361 | 81432576 | 2503 | 9235521 |
|  | 75 | 6120676 | 4982 | 28451556 | 9385 | 9916201 | 5542 | 844561 |
|  | 2549 | 550564 | 10316 | 27667600 | 6236 | 32661225 | 6461 | 25090081 |
|  | 1807 | 51984 | 5056 | 7414729 | 521 | 5080516 | 1452 | 9947716 |
|  | 1579 | 1750329 | 2333 | 110889 | 2775 | 1784896 | 4606 | 10432900 |
|  | 256 | 962361 | 2000 | 190969 | 1439 | 1153476 | 1376 | 5184 |
|  | 1237 | 226576 | 2437 | 2073600 | 2513 | 60025 | 1304 | 656100 |
|  | 1713 | 38912644 | 3877 | 13771521 | 2758 | 32205625 | 494 | 18913801 |
|  | 7951 | 50665924 | 166 | 2712609 | 8433 | 47128225 | 4843 | 18147600 |
|  | 833 | 27520516 | 1813 | 6724 | 1568 | 1142761 | 583 | 43322724 |
|  | 6079 | 27227524 | 1731 | 137569441 | 499 | 1038361 | 7165 | 45927729 |
|  | 861 | 339889 | 13460 | 172501956 | 1518 | 1177225 | 388 | 66912400 |
|  | 278 | 80586529 | 326 | 31899904 | 433 | 14784025 | 8568 | 13111641 |
|  | 9255 | 81396484 | 5974 | 15792676 | 4278 | 580644 | 4947 | 2965284 |
|  | 233 | 20484676 | 2000 | 622521 | 5040 | 5597956 | 3225 | 6579225 |
|  | 4759 | 12236004 | 1211 | 30891364 | 2674 | 1929321 | 660 | 31329 |
|  | 1261 | 274576 | 6769 | 2085136 | 4063 | 15460624 | 483 | 175561 |
|  | 737 | 1790244 | 5325 | 4713241 | 131 | 535824 | 64 | 286828096 |
|  | 2075 | 31329 | 3154 | 422500 | 863 | 53275401 | 17000 | 190357209 |
|  | 1898 | 96100 | 3804 | 1876900 | 8162 | 43811161 | 3203 | 126736 |
|  | 2208 | 4743684 | 2434 | 1249924 | 1543 | 208849 | 3559 | 8714304 |
|  | 30 | 91030681 | 1316 | 16386304 | 2000 | 1265625 | 607 | 1279161 |
|  | 9571 | 7263025 | 5364 | 11992369 | 875 | 717409 | 1738 | 28224 |
|  | 6876 | 19600 | 1901 | 75625 | 1722 | 73984 | 1570 | 33362176 |
|  | 6736 | 912025 | 2176 | 13689 | 1994 | 6105841 | 7346 | 25341156 |
|  | 5781 | 3617604 | 2293 | 13483584 | 4465 | 9840769 | 2312 | 21104836 |
|  | 7683 | 43007364 | 5965 | 29041321 | 1328 | 80910025 | 6906 | 1896129 |
|  | 1125 |   | 576 |   | 10323 |   | 8283 |   |
|  | **140337** | **865361634** | **147943** | **1235082645** | **160469** | **1556337207** | **184319** | **1284469066** |

**3.11.2 Relative Efficiency of Systematic against Simple Random Sampling**

The relative efficiency of systematic sampling as compared to simple random sampling in case of with and without replacement sampling is given as:

 

 

**3.12 CIRCULAR SYSTEMATIC SAMPLING**

It has been mentioned earlier that when *N* is not a multiple of *n* we get a biased estimator. We overcome this difficulty by adopting another scheme called CIRCULAR SYSTEMATIC SAMPLING. According to this technique we select a random start from 1 to *N* and therefore every *k*th unit is selected in a cyclical manner till required sample of size *n* is obtained. This procedure was suggested by Lahiri (1952). If *r* is a number selected at random from 1 to *N*, the sample consists of the units corresponding to the number:.

*r* + *ik* if *r* + *ik* ≤ *N*

and *r* + *ik* – *N* if *r* + *ik* > *N*

where i = 0, 1, 2, …., *n* – 1 and *k* being integer

It can be easily seen that probability of selection is equal for all units and is 1/*N*. The unbiasedness under circular systematic is shown below by an artificial example. Let *N* = 14, (1,2,3,…14) *n* = 5 and *k* = 3. All possible samples with different random starts are as:

|  |  |  |  |
| --- | --- | --- | --- |
| Random Start point  | Possible Samples | Sample Total | Sample mean |
|  | 1, 4, 7, 10, 13 | 35 | 7.0 |
|  | 2, 5, 8, 11, 14 | 40 | 8.0 |
|  | 3, 6, 9, 12, 1 | 31 | 6.2 |
|  | 4, 7, 10, 13, 2 | 36 | 7.2 |
|  | 5, 8, 11, 14, 3 | 41 | 8.2 |
|  | 6, 9, 12, 1, 4 | 32 | 6.4 |
|  | 7, 10, 13, 2, 5 | 37 | 5.4 |
|  | 8, 11, 14, 3, 6 | 42 | 8.4 |
|  | 9, 12, 1, 4, 7 | 33 | 6.6 |
|  | 10, 13, 2, 5, 8 | 38 | 7.6 |
|  | 11, 14, 3, 6, 9, | 43 | 8.6 |
|  | 12, 1, 4, 7, 10 | 34 | 6.8 |
|  | 13, 2, 5, 8, 11 | 39 | 7.8 |
|  | 14, 3, 6, 9, 12 | 44 | 8.8 |

  = 105.0  = 

Hence under circular selection procedure  even when n is not a multiple of N.

**exercises**

1. In a population with *N* = 6 and *K* = 2, the units *Yhi* are 2, 4, 6 in the first stratum and 8, 12, 16 in the second stratum. A sample of 4 units is to be taken as *n*1 = *n*2 = 2. Draw all possible samples and show that the sample mean is unbiased estimator of population mean. Find . Find *nh* under proportional and optimum allocation. Find the variance of sample mean for proportional and optimum allocations and compare with the variance of simple random sampling.
2. In a sample survey designed to estimate total number of cattle, the population of 2072 farms was divided into 5 strata by total average of the farms. A simple random sampling of farms was taken from each stratum and following information was recorded to estimate the total number of cattle in the population. Find the standard error of this estimate.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stratum Size | Number of Farms (*Nh*)  | *nh* | Total Number of Cattle (Σ*yhi*) | Σ*y*2*hi* |
| 0 – 15 | 635 | 153 | 619 | 5579 |
| 16 – 30 | 570 | 138 | 1423 | 24253 |
| 31 – 50 | 475 | 115 | 1758 | 34082 |
| 51 – 75 | 303 | 73 | 1691 | 51419 |
| 76 – 100 | 89 | 21 | 603 | 18305 |

1. A simple random sample of 10 villages from each stratum is drawn from the population divided into 3 strata regarding density of population (high, medium, low) and total number of households in the sample area are given below. Estimate the total number of household in the population and find standard error of your estimate.

|  |  |  |  |
| --- | --- | --- | --- |
| Stratum | *Nh* | *nh* | Number of Households |
| High | 510 | 10 | 84, 96, 87, 102, 99, 98, 90, 85, 90, 95 |
| Medium | 632 | 10 | 50, 40, 56, 47, 50, 53, 40, 41, 43, 46 |
| Low | 840 | 10 | 17, 25, 7, 0, 15, 7, 3, 0, 5, 15 |

1. For a socio-economic survey, all the villages in the region were grouped into four strata on the basis of the altitude above sea level and population density. From each stratum, 10 villages were selected using simple random sampling technique. The data on the number of households in each of the samples villages are given below:

|  |  |  |
| --- | --- | --- |
| Stratum | Total number of villages | Total number of households |
| 1 | 1411 | 43, 84, 98, 0, 10, 44, 0, 124, 13, 0 |
| 2 | 4705 | 50, 147, 62, 87, 84, 158, 170, 104, 56, 160 |
| 3 | 2558 | 228, 262, 110, 232, 139, 178, 334, 0, 63, 220 |
| 4 | 14997 | 17, 34, 25, 34, 36, 0, 25, 7, 15, 31 |

1. Obtain estimate of the total and calculate the standard error of this estimator.
2. Estimate the gain due to use of stratification as compared to un-stratified.
3. From the following population find *n*h under optimum and proportional allocation when n = 100. Also find the  when cost is given.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stratum | *Nh* | *S*2*h* |  | *Ch* |
| 1 | 300 | 196 | 60 | 25 |
| 2 | 200 | 81 | 50 | 36 |
| 3 | 500 | 64 | 30 | 16 |
| 4 | 150 | 36 | 20 | 9 |

1. 500 farms are divided into 4 strata as given below. The purpose is to estimate the total number of goats in 500 farms. A sample of 50 farms was selected taken with proportional allocation and number of goats in each farm and in each stratum are given. Estimate the total number of goats in that population and calculate the standard error of the total

|  |  |  |  |
| --- | --- | --- | --- |
| Stratum | *Nh* | *nh* | Number of goats in each farm |
| 1 | 40 | 4 | 76, 70, 75, 80 |
| 2 | 80 | 8 | 51, 45, 49, 45, 42, 50, 46, 43 |
| 3 | 15 | 15 | 30, 31, 35, 34, 33, 29, 25, 21, 31, 35, 37, 38, 39, 33, 35 |
| 4 | 230 | 23 | 10, 15, 18, 16, 17, 18, 10, 11, 8, 9, 10, 15, 9, 8, 13, 15, 16, 17, 12, 11, 9, 7, 8 |

1. Using the data given below, compare the efficiencies of following alternative allocation of a sample of 3000 factories for estimating the total output;
(i) proportional allocation, (ii) allocation proportional to total output,
(iii) optimum allocation.

|  |  |  |  |
| --- | --- | --- | --- |
| S.No. | No. of Factories | Output per factory | *Sh* |
| 1 | 18260 | 100 | 80 |
| 2 | 4315 | 250 | 200 |
| 3 | 2213 | 500 | 600 |
| 4 | 1057 | 1760 | 1900 |
| 5 | 567 | 2250 | 2500 |

1. Given the following information of 4 schools, estimate the proportion of students who visited the doctor at least once during the past year. *Ph* is known from the preliminary survey. Find .

|  |  |  |  |
| --- | --- | --- | --- |
| School | *Nh* | *nh* | *ph* |
| 1 | 2000 | 100 | 0.2 |
| 2 | 1600 | 80 | 0.3 |
| 3 | 1200 | 60 | 0.4 |
| 4 | 1200 | 60 | 0.3 |

1. The following information on literacy is available for an area. If a proportionate stratified sample is to be used in the near future for estimating of literate persons with the coefficient of 10%. Find the sample size needed. Compare the  and .

|  |  |  |
| --- | --- | --- |
| Age Group | Number of Persons | Proportion Literate |
| 15 – 24 | 25200 | 0.5 |
| 25 – 34 | 19100 | 0.3 |
| 35 – 49 | 36300 | 0.1 |
| 50 and over | 19400 | 0.01 |

1. The number of Pepper standards for selected villages in each of the three strata are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Stratum | Total number of Villages*Nh* | Villages Selected *nh* | Number of Pepper standards in each of the selected village |
| 1 | 441 | 11 | 41, 116, 19, 15, 144, 159, 212, 57, 28, 119, 76 |
| 2 | 405 | 12 | 39, 70, 38, 37, 161, 38, 27, 119, 36, 128, 30, 208 |
| 3 | 103 | 7 | 252, 385, 192, 296, 115, 159, 120 |

Estimate the total number of pepper standards also estimate the gain in precision due to stratification.

1. An investigator desires to take a stratified random sample with following assumption:

|  |  |  |  |
| --- | --- | --- | --- |
| Stratum | *Nh* | *Sh* | *Ch* in (Rs.) |
| 1 | 400 | 10 | 4 |
| 2 | 600 | 20 | 9 |

1. Estimate the values of *n*1/*n* and *n*2/*n* which minimize the total cost
*C* = *c*1*n*1 + *C*2*n*2 for given value of .
2. Estimate the total sample size required, under optimum allocation to make  where f.p.c. is ignored.
3. 2000 cultivators holding were stratified according to their sizes. The results are given as

|  |  |  |  |
| --- | --- | --- | --- |
| Stratum number | Number of holdings Nh | Mean are under wheat per-holding Yh | Shper-holding |
| 1 | 394 | 5.4 | 8.3 |
| 2 | 461 | 16.3 | 13.3 |
| 3 | 381 | 24.3 | 15.1 |
| 4 | 334 | 34.5 | 19.8 |
| 5 | 169 | 42.1 | 24.5 |
| 6 | 113 | 50.1 | 26.0 |
| 7 | 148 | 63.8 | 35.2 |

For a sample of 200 farms, compute the sample size in each stratum under proportional and optimum allocation. Calculate variance of estimated area under wheat form the sample and find relative precision as compared to simple random sampling.

1. If f.p.c. is ignored prove that
 
If the variance is fixed then for minimum cost the sample size for optimum allocation is

 

Also find the sample size if the cost per unit is the same.
2. With two strata, a sampler would like to have *n*1 = *n*2, instead of using the values given by Neyman allocation. If ,  denote the variance given by the *n*1 = *n*2 and the Neyman allocation, respectively. Show that the fractional increase in variance is

 

where r = n1/n2 given by Neyman allocation.

 **[**Hint: Since n1 = n2 and there are two strata

 

 and

 

 then

 

 where *r* = *W*1*S*1/*S*2*S*2**]**

1. A sampler has two strata with relative sizes *W*1 and *W*2. He believes that
*S*1 and *S*2 can be taken as equal. He would prefer to use proportional allocation but does not wish to incur a substantial increase in variance compared with optimum allocation. For a given cost *C* = *C*1*n*1 + *C*2*n*2, ignoring f.p.c. show that

 

 **[** Hint (i)  = 

 = 

 and n1 = , n2 =  as all Ss are equal

 Cost = , = 

 or  (1)

 (ii)  = 

  = 

 Hence  (2)

 From (1) and (2) we get required result**]**

1. Sample of 30 Mohallahs is drawn from a population of 210 Mohallas. The selection is systematic based on 1 in 7 Mohallah (k = 7). The following is the record of number of farm holdings of these Mohallahs. Estimate the total number of farm holdings in total Mohallahs. Under the assumption that the numbering is random find the variance of sample mean;

25, 30, 35, 37, 30, 40, 50, 35, 37, 40, 42, 45, 46, 40, 50, 38, 49, 45, 43, 35, 36, 39, 38, 25, 26, 40, 40, 51, 52, 46.

1. We have a natural population 1, 2, 3, 4, …, 50. Draw all possible samples (systematic) with k = 5. Find its mean and variance. Now from the above population draw a random sample of size 10, and find the mean and variance. Compare the variances of two selections and interpret the result.
2. We have a population of 5 units (1, 2, 3, 4, 5) and k = 2. If the random start is 1 then select the sample and if the random start is 2 also select the sample. Find, in each case, the unbiased estimator of population mean.
3. From the following population with N = 4. The size of the households is 5, 3, 3, 7, 4, 4, 6, 6, 4, 5, 3, 7, 7, 6, 4, 5, 6, 3, 5, 1, 3, 5, 6, 4, 4. Select a circular systematic sampling of size 5 households. Find the standard error under the assumption that the selection is random when
4. the random start is 14.
5. random start is 12.

 Compare the variances of the two samples.

1. From the data given in Example 4.2,
2. draw five circular systematic samples of size 7 each, from a rearranged frame.
3. from each of the five samples, estimate the total cultivated area in the tehsil.
4. obtain a single combined estimate from the five sample estimates. Also, calculate the standard error of this combined estimate.
5. Systematic sample of 29 plots has been selected from 290 plots. y-denotes the area (acres) under cultivation. Estimate the mean and variance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|   | y | Plot | y | Plot | y |
|  | 0.0 |  | 2.8 |  | 2.3 |
|  | 0.9 |  | 2.6 |  | 2.9 |
|  | 0.0 |  | 3.3 |  | 2.1 |
|  | 0.0 |  | 2.5 |  | 6.3 |
|  | 0.3 |  | 3.4 |  | 8.2 |
|  | 0.1 |  | 2.8 |  | 5.4 |
|  | 0.5 |  | 4.1 |  | 6.5 |
|  | 3.1 |  | 4.9 |  | 6.6 |
|  | 2.8 |  | 6.0 |  | 4.1 |
|  | 2.7 |  | 5.4 | -- | -- |

1. Following data shows the volume of timber of (y) for each strip:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | (y) |  | (y) |  | (y) |
|  | 762 |  | 471 |  | 165 |
|  | 651 |  | 426 |  | 224 |
|  | 461 |  | 448 |  | 192 |
|  | 521 |  | 402 |  | 161 |
|  | 653 |  | 372 |  | 104 |
|  | 544 |  | 372 |  | 94 |
|  | 542 |  | 411 |  | 102 |
|  | 59 |  | 323 |  | 115 |
|  | 533 |  | 381 |  | 110 |
|  | 517 |  | 430 |  | 109 |
|  | 520 |  | 434 |  | 83 |
|  | 539 |  | 321 |  | 36 |
|  | 509 |  | 543 |  | 61 |
|  | 449 |  | 607 |  | 92 |
|  | 492 |  | 416 |  | 75 |
|  | 498 |  | 326 |  | 64 |

1. Examine the behaviour of the sampling variance of estimates of volume of timber based on systematic samples of sizes 4, 8 and 12.
2. Compare the efficiency of systematic sampling with those of simple random sampling without replacement for the sample sizes considered in (i).
3. Also study the efficiency of sampling the strips with probability proportional to the length of the strips with replacement.
4. Given below are data for 10 systematic samples of size 4 from the population of 40 units.

|  |
| --- |
| Systematic Sample Numbers |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 0 | 1 | 3 | 4 | 4 | 5 | 6 | 7 | 7 | 9 |
| 7 | 8 | 9 | 10 | 12 | 13 | 15 | 6 | 16 | 17 |
| 18 | 18 | 19 | 20 | 21 | 20 | 24 | 13 | 28 | 29 |
| 29 | 30 | 31 | 31 | 33 | 32 | 35 | 37 | 38 | 63 |

Work out the relative efficiency of systematic sampling over simple random sampling.

1. Following data relating to area under guava crop in some district of Punjab. Draw a sample of size 5; (i) using simple random sampling method, (ii) using systematic sampling method; and compare the efficiency of these two methods.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No. | Area under guava crop | S.No. | Area under guava crop | S.No. | Area under guava crop |
|  | 166.15 |  | 10.34 |  | 3.90 |
|  | 24.72 |  | 95.16 |  | 15.31 |
|  | 100.77 |  | 22.40 |  | 1.44 |
|  | 87.14 |  | 10.97 |  | 14.88 |
|  | 116.28 |  | 39.07 |  | 23.01 |
|  | 60.22 |  | 13.70 |  | 3.44 |
|  | 13.59 |  | 26.64 |  | 14.32 |
|  | 41.70 |  | 1.40 |  | 24.39 |
|  | 10.52 |  | 12.57 |  | 9.88 |
|  | 13.85 |  | 2.00 |  | 17.66 |
|  | 12.92 |  | 6.72 |  | 3.26 |
|  | 10.73 |  | 20.75 |  | 6.89 |
|  | 38.64 |  | 51.65 |  | 0.84 |
|  | 15.92 |  | 16.42 |  | 13.02 |
|  | 9.09 |  | 3.90 |  | 32.85 |
|  | 155.51 |  | 2.44 |  |  |

1. Following table furnishes complete enumeration data on length (x) of strip and volume (y) of timber for each strip in 3 blocks of the Block Mountain Forest, California.

|  |  |  |
| --- | --- | --- |
| Block No. I | Block No. II | Block No. III |
| Strip No. | x | y | Strip No. | x | y | Strip No. | x | y |
|  | 12 | 762 |  | 9 | 471 |  | 6 | 165 |
|  | 12 | 651 |  | 9 | 426 |  | 6 | 224 |
|  | 12 | 461 |  | 9 | 448 |  | 6 | 192 |
|  | 12 | 521 |  | 9 | 402 |  | 6 | 161 |
|  | 12 | 653 |  | 9 | 372 |  | 6 | 104 |
|  | 12 | 544 |  | 9 | 372 |  | 5 | 94 |
|  | 12 | 542 |  | 9 | 411 |  | 5 | 102 |
|  | 12 | 590 |  | 9 | 323 |  | 5 | 115 |
|  | 11 | 533 |  | 9 | 381 |  | 4 | 110 |
|  | 11 | 517 |  | 9 | 430 |  | 4 | 109 |
|  | 11 | 520 |  | 9 | 434 |  | 4 | 83 |
|  | 11 | 539 |  | 9 | 324 |  | 4 | 36 |
|  | 10 | 509 |  | 9 | 543 |  | 4 | 61 |
|  | 10 | 449 |  | 9 | 607 |  | 4 | 92 |
|  | 10 | 492 |  | 8 | 416 |  | 4 | 75 |
|  | 10 | 498 |  | 8 | 326 |  | 4 | 64 |

 Source:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Examine the behaviour of the sampling variance of estimates of volume of timber based on systematic samples of sizes 2, 6 and 12.
2. Compare the efficiency of systematic sampling with those of simple random sampling with and without replacement for the sample sizes considered in (i).
3. Also study the efficiency of sampling the strips with probability proportional to the length of the strips with replacement.
4. A list of 108 villages in a Tehsil arranged in ascending order of geographical area (x) is given as together with village-wise area under winter paddy (y).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. | x | y | No. | x | y | No. | x | y |
|  | 12 | 7 |  | 264 | 102 |  | 515 | 272 |
|  | 106 | 33 |  | 264 | 102 |  | 541 | 155 |
|  | 120 | 87 |  | 266 | 187 |  | 542 | 292 |
|  | 120 | 78 |  | 271 | 23 |  | 543 | 214 |
|  | 121 | 56 |  | 273 | 129 |  | 562 | 275 |
|  | 121 | 62 |  | 274 | 51 |  | 570 | 100 |
|  | 124 | 58 |  | 280 | 161 |  | 586 | 418 |
|  | 128 | 19 |  | 287 | 179 |  | 601 | 189 |
|  | 135 | 64 |  | 292 | 76 |  | 653 | 129 |
|  | 137 | 61 |  | 313 | 137 |  | 658 | 230 |
|  | 145 | 74 |  | 320 | 127 |  | 678 | 396 |
|  | 147 | 13 |  | 324 | 104 |  | 681 | 289 |
|  | 151 | 81 |  | 327 | 115 |  | 682 | 166 |
|  | 153 | 41 |  | 333 | 106 |  | 691 | 83 |
|  | 160 | 58 |  | 349 | 245 |  | 698 | 232 |
|  | 166 | 44 |  | 350 | 117 |  | 710 | 282 |
|  | 176 | 65 |  | 364 | 170 |  | 716 | 191 |
|  | 178 | 69 |  | 365 | 210 |  | 716 | 305 |
|  | 185 | 29 |  | 370 | 98 |  | 727 | 303 |
|  | 206 | 46 |  | 379 | 270 |  | 730 | 288 |
|  | 209 | 93 |  | 389 | 79 |  | 738 | 286 |
|  | 216 | 38 |  | 396 | 99 |  | 805 | 239 |
|  | 224 | 87 |  | 397 | 147 |  | 808 | 242 |
|  | 229 | 72 |  | 400 | 187 |  | 864 | 146 |
|  | 230 | 127 |  | 404 | 273 |  | 873 | 445 |
|  | 236 | 114 |  | 410 | 118 |  | 897 | 487 |
|  | 238 | 88 |  | 418 | 130 |  | 910 | 354 |
|  | 240 | 108 |  | 433 | 158 |  | 924 | 340 |
|  | 241 | 94 |  | 446 | 116 |  | 1034 | 401 |
|  | 243 | 116 |  | 453 | 194 |  | 1117 | 261 |
|  | 244 | 58 |  | 460 | 161 |  | 1156 | 613 |
|  | 246 | 47 |  | 462 | 222 |  | 1196 | 227 |
|  | 248 | 69 |  | 467 | 223 |  | 1323 | 704 |
|  | 249 | 44 |  | 501 | 96 |  | 1419 | 682 |
|  | 251 | 56 |  | 503 | 164 |  | 1473 | 373 |
|  | 259 | 160 |  | 514 | 318 |  | 1496 | 164 |

 (x and y in acres; 1 acre = 0.4047 hectare).

1. Draw 4 circular systematic samples of 7 villages each with the following five independent random starts: 45, 3, 18, 62 and 37.
2. Making use of the 35 sample observations obtained in (i), estimate the relative efficiency of systematic samples as compared to that of simple random sampling without replacement for estimating the total area under peddy (Y) based on a sample of 7 villages.
3. Obtain a single combined estimate of Y based on all the 5 samples drawn in (i) and also estimate its use.

**Table 3.2**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sr.No. | Population (000) | Cultivated Area (Acres) | Sr.No. | Population | Cultivated Area (Acres) |
|  | 226 | 678 |  | 904 | 700 |
|  | 670 | 663 |  | 773 | 602 |
|  | 4505 | 1290 |  | 1040 | 532 |
|  | 1732 | 1170 |  | 760 | 438 |
|  | 2874 | 1390 |  | 2084 | 633 |
|  | 2282 | 1110 |  | 828 | 277 |
|  | 793 | 760 |  | 4877 | 1640 |
|  | 895 | 730 |  | 911 | 424 |
|  | 1157 | 950 |  | 1205 | 822 |
|  | 3201 | 1700 |  | 1139 | 555 |
|  | 1117 | 909 |  | 4064 | 347 |
|  | 1236 | 1169 |  | 1114 | 744 |
|  | 5201 | 1840 |  | 547 | 372 |
|  | 848 | 660 |  | 1175 | 644 |
|  | 1238 | 1140 |  | 1159 | 732 |
|  | 1917 | 1360 |  | 441 | 622 |
|  | 1800 | 1509 |  | 555 | 342 |
|  | 2335 | 1810 |  | 827 | 387 |
|  | 4396 | 2240 |  | 2867 | 322 |
|  | 1607 | 1225 |  | 726 | 636 |
|  | 2071 | 1250 |  | 633 | 410 |
|  | 2166 | 1690 |  | 680 | 427 |
|  | 7780 | 3200 |  | 587 | 496 |
|  | 2746 | 1744 |  | 1901 | 936 |
|  | 2549 | 2400 |  | 2419 | 1226 |
|  | 1007 | 680 |  | 1258 | 836 |
|  | 1567 | 970 |  | 1225 | 634 |
|  | 5271 | 1850 |  | 1447 | 978 |
|  | 659 | 340 |  | 1314 | 724 |
|  | 3209 | 2450 |  | 1298 | 422 |
|  | 2902 | 1760 |  | 728 | 493 |
|  | 2955 | 2120 |  | 851 | 396 |
|  | 1746 | 1220 |  | 786 | 732 |
|  | 1045 | 860 |  | 663 | 422 |
|  | 666 | 620 |  | 740 | 370 |