***CHAPTER- 1***

**CONCEPT OF SAMPLING**

**1.1 Introduction:**

Most survey work involves sampling from finite populations. There are two parts to any sampling strategy. First there is the selection procedure, the manner in which sampling units are to be selected from a finite population. Second there is the estimation procedure, which prescribes how inferences are to be made from sample to population. These inferences may be either enumerative or analytical. Enumerative inference seeks only to describe the finite population under study whereas analytical inference attempts to explain the underlying distributional and functional characteristics of a population. Enumerative inference typically concerns with the estimation of some parameters of a population such as means, totals, proportion and ratios. Viewing the same population analytically we might seek to regress household income on such variables as number of employed adults, educational level of household head etc. regression implies an explanatory model.

Analytical inference consists of specifying appropriates of a model adequately describes the sample, and hence the population from which it was selected. It is customary to distinguish between enumerative and analytical inference in terms of complexity of the population characteristics estimator.

Estimating a mean, total etc. is regarded as an enumerative problem

whereas

estimating a regression or correlation coefficient as an analytical one,

but if the mean in question is a parameter of a simple explanatory model, the inference, the inference is enumerative.

Analytical and enumerative inference the model used provides its own probability structure, in-terms of which the inference can proceed. Same methodology is used in every other area of statistics.

**For enumerative inference, a quite different probability structure is used;** it depends on the manner in which the sample is selected. This is the classical finite population sampling inference developed by Neyman (1934), who based his results on Gram (1883) and Bowley (1913).

There is hardly any field of statistics that has attracted so much attention in post-war period as the theory of sampling. This was probably because of its many practical applications. It is sufficient to recall that the use of sampling theory has entirely changed the work of data collection and seems to be the reason why most of the names among contemporary statisticians have devoted much of their time to some problems in the theory and practices of sample surveys.

Systematic interest in the use of sampling theory appeared towards the end of the last century when Kiaer (1890) used the representative method for collecting data independently of the census. In 1901 he also demonstrated empirically that stratification could provide good estimates of finite population totals and means. On the recommendation Kiaer International Statistical Institution 1903 recommended the adoption of stratified sampling with proportional allocation as an acceptable method of data collection. In line with the present thinking, the selection procedure used by Kiaer and others dealt with a method of drawing and representative sample from the population. The primary difference in the earlier uses of the sampling methods lay in the selection of the sample, it was brought that a purposive selection based on sampler’s knowledge of the population with regard to closely correlated characters was the best way of getting a sample that could be considered representative of the population. Neyman’s (1934) was the revolutionary paper. In that paper he made it clear that random selection had its basis on a sound scientific theory, which definitely gave to sample survey the character of a n objective research tool and made it possible to predict the validity of the survey estimates. Actually in this sense it can be said to mark the beginning of a new era. He suggested optimum allocation in stratified sampling. Now a days sampling theory has been so developed in its application that it is widely used in all the fields of life and the use of this theory is made fruitful in the underdeveloped areas.

**2.1 BASIC DEFINITIONS**

SAMPLING is primarily concerned with the collection of data of some selected units. It is also a procedure or process of selecting some units out of the whole lot of the individuals with some common characteristics; i.e. population of persons, farms, houses, etc. A population is said to be an INFINITE population if the number of units are finite. The whole lot or aggregate of individual sampling involves making inferences about the population on the basis of observations of some units from the population. Almost all of the modern developments in statistics relate to the inferences that can be made about a population when information is available from only a part of the population. CENSUS is another method of data collection and is a complete enumeration of the population, for instance census of population, census of agriculture, census of manufacturing industries etc. An elementary unit or simply a unit is an identifiable element of the population on which information be made, e.g. person, family, farm, etc. The sampling units selected in the sample may be termed sample units and the values of the characteristics under study for the sample units are known as sample observations. A list of elementary units from which a sample is selected is called the SAMPLING FRAME. This may be a list of population units with identification or maps showing the boundaries from where sample is to be selected. The characteristics of the frame should be that it must be up-to-date and free from any error. The number of sampling units selected in the sample is termed sample size.

**1.2.1Types of Sampling**

Generally samples are of two types;

1. A Probability or Random Sample is one if its selection follows by certainable laws of chance. The probability of selection of each unit may or may not be independent of the selection of units. Infect this is a general name given to the sampling plan in which:
   1. Every member of the sampled population has a known probability of being included in the sample,
   2. The sample is drawn by some random number table randomization process in consonance with the probabilities and
   3. In the analysis of the sample, weights appropriate to the probabilities are used.

It is not necessary that the probability of selection should be equal for every member of population.

1. A sample selected by a non-random process is termed as non-probability or nonrandom sample, judgment sample or purposive sample. This type of sample design in useful when the population is highly variable and the sample is small. This type of sampling is seldom used in large-scale surveys. In non-probability sampling there is no way out to check the accuracy and to obtain the accurate estimates.

**1.2.2Sampling and Non-sampling error**

Since sample is a part of population, the result based on the observations in the sample will not correspond to that of population, there must be some difference and this difference is know as SAMPLIG ERROR. In other words the error arising due to drawing inferences about the population on the basis of sample observations is termed as sampling error. The sampling error usually decreases as the sample size increases and, infect, in many situations the decrease is inversely proportional to the square root of the sample size (see Yates 1949). The behavior of sampling of sampling error with sample size can be seen is Fig. 1.1.

A sampling error is reduced to minimum if the selection procedure, choice of sampling unit, size is the most important factor, any error other than sampling error is termed as non-sampling error which includes:

1. Measurement error
2. Response error
3. Recording error
4. Coding error
5. Use of improper estimates
6. Non-response error etc.

**1.2 Bias and Mean Square Error**

Bias is the difference between the expected value if an estimate and the population value. If y is an estimator and Y is the population value and then the bias may be defined as:

bias = E(y) – Y

If E(y)>Y then the bias is positive and if E(y)<Y then the bias is negative. A positive bias implies that sampling plan gives high estimates and negative bias means that the estimates are low. Mean Square Error is a measure of the extent to which an estimated value differs from the true population value. If the estimates are unbiased then the mean square error (M.S.E.) is equivalent to the variance.

M.S.E. = Variance +(Bias)2

**1.2.4 Advantages of sampling**

The advantages of sampling over complete enumeration have been outlined below:

1. As sample is a part of population, the information can be collected cheaper and more rapidly as compared to complete enumeration.
2. A sample makes it possible to concentrate on individual units and to obtain relevant information comprehensively
3. Non-sampling error can be reduced, by adopting a suitable sample design.
4. More precise results can be obtained by employing persons of high quality.

**1.2.5Randomness:**

In this book we shall make free use of the word random. A unit is said to be random if it has the same chance of selection in the sample. Suppose in a deck of 52 cards, we draw a black card. I shall be random if all the black cards have the same chance of being drawn, i.e. if all the black cards are equally likely to be drawn, the process of drawing one card shall be a random process.

In sampling, we select a few observations from the population. The method of choosing a sample is very important as it is supposed to be representative of the result based on it shall be generalized to the population. If some observations are more likely to be selected than others, then the sample drawn will be biased and the results based on it shall not be reliable. In order to present this type or any-other type of bias, Fisher (19 ) has suggested the use of random process of selection draw a sample. Random number tables can be used. Suppose we have a population of 100 observations coded by its number or name. The slips are thoroughly mixed up in a box. A slip is drawn and its number noted and is replaced in a box before the second slip is drawn. Suppose a sample of size 20 is needed than we shall draw 20 slips. These 20 slops will be our sample. An observation can be repeated as many times as it is drawn. If we decide to have 20 distinct observations, then the repeat number is rejected and replaced in the box before the next slip is drawn. This procedure is clumsy particularly in a large sample. We can make use of random number tables for selecting a proper sample. The procedure of drawing a random sample is given in Fisher and Yates (19 ) table.

Generally speaking samples are of two types:

1. probability sample or Random sample is one in which the probability of selection of each unit in the population is known as non-zero chance of being selected in the sample. The probability of selection of each init may or may or may not be independent. Infect this is a general name given to the sampling plan given be Cochran, Mosteller and Theckey, 1954 in which “(a) every individual in the sampled population has a known probability of entering in the sample, (b) the sample is chosen by a process involving one or more steps of automatic randomization, (c) in the analysis of the samples, weights appropriates to the probabilities (a) are used”. Valid inference form the sample can be derived by using the theory of probability
2. A sample selected by a non-random process is termed as non-probability sample e.g. Judgment Sample or Purposive Sample. This type of sampling design is useful when the population is highly variable and the sample is small. In non-probability sample there is no way our to check the accuracy and to obtain the accurate estimates.

We are unable to determine the sampling error as in the non-probability sample, the probably that a specified unit is selected cannot be calculated and the frequency distribution of the estimates produced by this process cannot be determine.

The work statistics is derived from the Latin word ‘Status’ meaning a state. Statistics means the affairs of a state. This definition was in use much before the Roman Empire, but when Caesar Agustus issued a decree that every body living in his territory must pay tax to the state, every individual had to register himself with the nearest tax collector. By this registration system, the ruler was in a position to count the total population, which was considered strength of the state as the total count enabled the state to enlist the persons able to fight. As such, statis was termed as the science of Kings. Later on the registration of wealth, taxes, marriages, divorces, deaths, births etc. were also instituted, in Indo-Pak subcontinent, the first known use of statistics and its methods was made during the reign of Morya-Dynesty when a decree had been promulgated to register deaths and births. During the period when Shershah Suri by looking at the employment statistics realized that there was a large number of un-employed persons in his country, he decided to construct a road now known as Grand Trunk Road connecting Peshawar with Colcutta. Later he developed a technique of leaving revenue taxes on the persons who are capable of paying. This method was systematically taken up by the ruler, Akbar. Many statistical surveys were conducted in his reign

The earliest known big count of population was in 3050 B.C. in Egypt when Egyptians decided to construct Pyramids. Another population count was conducted in Egypt in 1400 B.C. and these figures were used to organize and train persons for war-fares, to collect taxes and to distribute land to peasants. The basic modern vital statistics is due to John Graun who studied birth and death statistics in England in the 17th century. At the end of that century, Casper Newman studied deaths, births, and marriages statistics, and Edmund Hally and Willium Potty constructed life tables on the basis of death and births statistics. Later on, life tables were used in life insurance business.

The sampling methods were discussed by Graunt (1662), Sussmilch (1741), Messance(1765), Mohean (1778), and Haplace (1786) in the seventeenth or eighteenth centuries. Messance (1765) and Mohean (1778) estimated the population of France and Laglace (1778) computed error in the population of France. Later, Eden (1800) estimated the Britain’s Population on the sampling dasis and the first census of Britain in 1801 confirmed his sample results. Kaer (1895) collected sample information on Norwegian workers and tabulated Norway census data in 1900 Denmark data in 1901 and Oslo data in 1913. Booth et al (1889-1891), Yule (1895) and Durkheim (1897) did some work. Japan also conducted sample enquiries in 1921 or 1923. At this stage, systematic work on sampling was developed. Some leading workers are Bowlen(1901), Tice(1929), Boring(1929), King (1930), Leven (1932), Neyman (1934), Yats (1949), Chevery (1949)

and many others in recent years.

Planning The Survey

If meaningful results are to be obtained from the surveys, it is essential that plan in respect of area of investigation, nature of survey, methods of enumeration and estimation, the size and design of surveys are carefully drawn well in advance of the actual execution of surveys. When planning surveys, it is necessary to study the needs for the data in consultation with its users and to ensure that the surveys are successfully completed within the available resources including time and money. In developing countries the most important factors in taking censuses and surveys are the cost and the trained personnel and the quality and size of survey vary in accordance with the objectives of surveys. The operations to be carried out during the planning period are also different for different types of surveys. Apart from this, certain basic principles are common to all types of surveys.

Some of the operations may need closer attention in one survey than the other. Detailed planning will have to be undertaken for each type of survey. A comprehensive discussion has been given by Yates (1965), however the following steps may be considered at the planning stage:

1. Objectives of the survey.
2. Specifications of population to be sampled.
3. Data to be colleted and methods of its collection
4. Questionnaire and its design.
5. The frame.
6. Sampling units
7. Sampling and non-sampling errors
8. Sample design
9. Selection of the sample
10. Pre and Post enumeration surveys.

The sampling unit and sampling and non-sampling errors are not discussed in this chapter separately.

Objectives of the Survey:

The most carefully planned surveys need clear out statements regarding their objectives. It becomes very difficult to properly work out of the details of operations without elucidation of its objectives. Unless the objectives of the surveys are known, it is not be possible to define population, sampling unit or other terms. For example, the major goals for the U.S. 1960 Census of Population and Housing were laid down as follows:

1. To improve the quality of the statistics.
2. To reduce the time between taking the censuses and publishing the results.
3. To achieve the needed results at relatively low cost per person and housing unit enumerated.

Likewise, the objectives of the Pakistan Sample Census of Agriculture (1960), had been defined as follows:

1. To develop basic information on the structure of Agriculture in Pakistan.
2. To provide detailed basic information about the national resources of agriculture and their state of use, for the sake of development planning, and
3. To provide benchmarks for the improvements of current agricultural statistics.

The Frame:

A frame is the strong hold of the design. In most of the developing countries, the proper frames are not available, but are constructed during the operation of surveys. For such reasons, a design can be simple if good frames are available and it can be complex if good frame are not available. Population is divided into units which are called sampling units. The list of units is defined as frame. Frame must be complete, accurate and efficient.

Advantages of Sampling

Besides using a fraction of population, it has many built in advantages:

* + - 1. It widens the scope of a study by investigating several aspects of a problem. In census, one can ask a few questions. Whereas in sampling, one can ask several questions.
      2. It enables statisticians to:
         1. Test various design for a sample survey and chose the best one,
         2. Test questionnaires and its performances for selecting the best one,
         3. Test investigators and chose the most efficient ones and possible effects of difference between investigators,
         4. Measure the extent of under-coverage or over-coverage of units,
         5. Widen the scope of the study by selecting several information on a number of units,
         6. Reduce expenditure,
         7. Measure survey errors, bias and computing errors.

Estimation:

In most of statistical studies, the population parameters like ‘average age of persons’, ‘average potency of a drug’, ‘variation in income of persons’ etc. are unknown. It is almost impossible to contact each and every person of a city for his age in order to determine an average age of persons of the city. For this we draw a small random sample of persons from the city and record their ages. The average age of samples persons is an estimate of the average age of persons of the city. If the sample is representative of the city, the average age will be close to the average age of persons.

The sample values of the population parameters are known as estimates. The estimates can be either a single value called point estimates or a range of values called estimating interval estimates.

Suppose a city has a population of 25,000 persons. We select a random sample of 1,000 persons and ask their ages. Suppose the average age of 1,000 persons is 26 years. Then 26 are a point estimate of average age of 25,000 persons. An interval estimate may specify the range of value say 24 to 28 indicating that we think that the average age of 25,000 persons based on the values of 1,000 persons lies between 24 years and 28 years.

The choice of and appropriate point estimate usually depends on how well the estimate satisfies certain criteria. There are 4 properties of a good estimate, viz

1. The property of unbiased-ness
2. The property of efficiency
3. The property of sufficiency
4. The property of consistency.

We shall discuss (i) only.

Unbiasedness:

We call a sample value an unbiasedness estimate of population value, if the average of the sample values equal the population.

Suppose a population consists of 4 units; (4, 5, 7, 0) then mean of 4 values,



Suppose we draw a sample of 2 units. If the mean of 2 units is unbiasedness of the mean of means of all possible sample of size 2 equals the population mean. Let us draw all possible samples of size 2 from 4 units:

**Table**

|  |  |  |
| --- | --- | --- |
|  | **Sample** | **Means** |
| 1. | (4, 5) |  |
| 2. | (4, 7) |  |
| 3. | (4, 0) |  |
| 4. | (5, 7) |  |
| 5. | (5, 0) |  |
| 6. | (7, 0) |  |
|  |  |  |
|  | Total | 24.0 |

Mean = 

This equals the population means of 4.0. It implies that the sample mean is an unbiased estimate of population mean.

Similar results can be obtained for variance.

Mathematical Preliminaries:

* + 1. **Random Variable**

If an experiment consists of observing the value of some variable X which can take a number of values xi (i = 1, 2, 3, . . . .,k) each with an associated probability Pr(X = xi) then X is a random variable.

* + 1. **Expectation:**

The expectation of a random variable X, taking the values xi (i = 1, 2, 3, . . . .,k) with probability pi (i = 1, 2, 3, . . . .,k), pi = 1 ia defined as

E(X) = xi pi (1.5.1)

The summation extends over the entire range of X the following results may be easily established

(i) E(cX) = c(X) (1.5.2)

where c is constant.

(ii) E(X1 +X2) = E(X1) + E(X2) (1.5.3)

where X1 and X2 are two random variables

E(c1X1 + c2X2 + ..+ckXk) = c1E(X1) + c2E(X2) +..+ ckE(Xk) (1.5.4)

* + 1. **Mean Square Error and Bias**

The mean square error about a given constant a of X is defined as

E(X – a)2 (1.5.5)

Which attains a minimum when taken about the mean µ and is called variance of X

Var(X) = E(X – µ)2 (1.5.6)

Note that

E(X – a)2 = E(X – µ + µ)2 = E(X – µ)2 + (X – µ)2

= Var(X) + (Bias)2 ≥ Var (X) (1.5.7)

Bias is the difference between the expected values of an estimate and the population value. If y (sample mean) is an estimator and Y is the population mean then the bias may be defined as :

Bias =

If  then the bias is positive and if  then the bias is negative. A positive bias implies that sampling plan gives high estimates and negativebias means that estimates are low.

* + 1. **Conditional Expectation and Variance**

Consider a pair of random variable X and Y. The conditional expectation of Y given X is defined as a function of S with value at X = x and written as

E(Y | x)

Where E refers to expectation with respect to Y. When a further expectation is taken with respect to x, we can find a general property of conditional expectation.

E(y) = ExEy (y|x) (1.5.8)

We can define the conditional variance of y given X = x

Var(y) = E(y2|x) – [E(y|x)]2 (1.5.9)

And obtain a general result of conditional variance,

Var(y) = Ex [Vary(y|x)] + Varx [E(y|x)]

= E1V2(y) + V1[E2(V)] (1.5.10)

which decompose the total variance of y as the sum of two components (i) the average conditional variance and (ii) the variance of the conditional average.

*CHAPTER 2*

SIMPLE RANDOM SAMPLING

**2.1 INTRODUCTION**

Random Sampling or more precisely, Simple Random Sampling is a term covering two of the most straightforward selection procedures used in probability sampling. In both these selection procedures population units are drawn with equal probabilities. If a unit once selected is not allowed to be selected again, the procedure is known as simple random sampling without replacement (Srswor). If the selection at each draw is from the whole population, the procedure is known as simple random sampling with replacement (Srswr). The selection of units using simple random sampling with replacement is independent from draw to draw and the selection of units using simple random sampling without replacement is not independent. This is because in simple random sampling without replacement probability of selection of a population unit at any given draw depends on whether or not another unit has been selected at some previous draw. **This selection procedure of simple random sampling with replacement is also known as unrestricted random sampling.**

**THEOREM (2.1)**

In simple random sampling with and without replacement the probability that any given population unit is selected at any particular draw is 1/N, where N denotes the number of units in the population. **◊**

**PROOF**

For simple random sampling with replacement the theorem follows immediately from the independence and identical nature of a draw. If first unit is selected, its probability of selection is . The unit is replaced before the second unit is drawn. In the second draw, the total population units is N. The probability of second unit is again . Thus for each draw the probability of its selection is .

For simple random sampling without replacement it is necessary to consider each draw individually.

Clearly each population unit has the probability 1/N of being selected at the first draw. To select a unit at the second draw, however, it is necessary that unit is not selected at the first draw. The probability that this unit is not selected at the first draw is . The probability of selection at the second draw conditional on its non-selection at the first draw is 1/(N-1), so the unconditional probability of selection at the second draw is the multiplication of probability that is not selected at the first draw and is selected at the second draw . This argument can be easily generalized to the kth draw, for which the unconditional selection probability is:

.**◊**

**ALTERNATIVE PROOF**

This may be proved by another way. Suppose that the population unit say ith is selected at the first draw, the probability of this unit will be 1/N. The conditional probability that the second selection will produce jth (j ≠ i) unit is 1/(N-1) as there will be N-1 units in the population. The chance of getting the jth unit at the first selection is 1/N. Hence the absolute probability that ith unit is included in the sample at the second draw is “the sum of the product that the jth unit is selected at the first and ith unit is selected at the second draw” i.e.

. **◊**

This result may be illustrated by a simple example. Consider a population of three units; A, B and C. Ignoring the order of selection, there are three possible samples of two distinct units. These may be denoted by {A B}, {B C} and {C A}. Each of them may be selected in two ways. Thus {A B} may be achieved by selecting A at the first draw and B at the second, or B at the first draw and A at the second. If sampling is without replacement, the probability of selecting A at the first draw is 1/3, and the conditional probability of selecting B at the second draw given that A was selected at the first is 1/2. Thus the probability of selecting A first and then B is 1/6. Similarly the probability of selecting B first and then A is 1/6. Hence the probability of selecting the sample {A B} is   
. The same is true for {B C} and {A C}.

A number of corollaries may be drawn from Theorem (2.1).

**COROLLARY (2.1.1)**

Suppose simple random sample of size n is drawn without replacement from a population of N units. The probability that any specified unit is in the sample is n/N. This may be proved by simple argument. The probability that the ith unit may either is in the sample at the first draw, or at the second draw, or at the third draw, and so on is 1/N. There are n possible draws, so that probability of inclusion of the ith unit to be in the sample is the sum of the probabilities that it is selected at the first or at the second or at the third draw and so on i.e. **.**◊**

Note: This ratio, n/N, plays an important role in simple random sampling theory. It is known as the *sampling fraction* and denoted by f.

This probability is generally called the probability of inclusion of the ith unit to be in the sample

**COROLLARY (2.1.2)**

Suppose a Simple Random Sampling of size n is drawn without replacement from a population of N. The member of possible samples using simple random sampling without replacement is . The proof is trivial.

**COROLLARY (2.1.3)**

Suppose a Simple Random Sampling of size n is drawn without replacement from a population of N. The probability of selection of a sample of size n is Each of these  possible samples occurs with the same probability i.e. . This may be proved as:

**PROOF**

The probability that a specified unit is in the sample is n/N, and the probability that another specified unit from the remaining N-1 units is in the sample is (n-1)/(N-1). Hence the probability that any particular sample  of size n is selected from a population of N units is

 =**◊**

**COROLLARY (2.1.4)**

The number of times the ith (or any other) population unit appears in the set of all these possible samples is .

The last three of these corollaries do not have any direct analogues in simple random sampling with replacement, because of the possibility of multiple selections. Consider for instance the nine possible outcomes of two units selected with replacement from the population of three units A, B, C. These are

AA, BA, CA, AB, BB, CB, AC, BC, CC,

where, for instance, CA denotes the outcome ‘C selected first, then A’. With simple random sampling with replacement each of these nine outcomes is equally probable, but the sample **{A B} can be formed in two ways and thus has twice the probability of occurrence of A A (the only outcome in which the unit A is selected twice).**

It is nevertheless possible to formulate corollaries for simple random sampling with replacement in terms of (ordered) outcomes rather than in terms of (unordered) samples.

**COROLLARY (2.1.5)**

There are Nn possible outcomes using simple random sampling with replacement (srswr).

**COROLLARY (2.1.6)**

Each of these outcomes occurs with the same probability.

**COROLLARY (2.1.7)**

The number of times the ith population unit appears in the set of all possible outcomes is nNn/N or nNn-1.

There are more possible samples using simple random sampling with replacement than there are using simple random sampling without replacement. Suppose, for instance, that we have a population of three units, A, B and C, and we select a sample of two units from it. Using simple random sampling without replacement there are six possible samples, which are equally likely:

A selected at the first draw, B at the second draw

A .. .. , C .. ;

B .. .. , A .. ;

B .. .. , C .. ;

C .. .. , A .. ;

C .. .. , B .. ;

We therefore, say there are six possible *ordered samples* AB, AC, BA, BC, CA and CB. But if we ignore the order in which the sample units are selected, the ordered samples AB and BA can be put together as the *unordered sample* AB, and so on. There are, then, three unordered samples, all equally likely, namely AB, AC and BC. It is a common practice to ignore orderings and to describe the unordered sample simply as *sample*. Since the sample AB, AC and BC are equally likely; each has a probability of selection of one third.

With simple random sampling with replacement (srswr), however, there are nine possible unordered samples, all equally likely, namely AA, AB, AC, BA, BB, BC, CA, CB and CC. If ordering is ignored, we have six possible unordered samples, AA, AB, AC, BB, BC and CC, but now the samples containing repeated selections (AA, BB and CC) are made up of only a single ordered sample while the others (AB, AC and BC) are each made up of two ordered samples, and we have twice the probability of selection. AA, BB and CC have a probability of selection equal to *one ninth*, while AB, BC, and CA each has a probability of selection equal to *two ninths*. Further, it is a common practice to ignore repeats (as well as to ignore orderings so the six samples possible under simple random sampling with replacement would be written as AB, AC and BC (each with probability two ninths) and AA, BB and CC (each with probability one ninth).

More generally, when n sample units are selected from N population units, simple random sampling without replacement yields



equally likely ordered samples and



equally likely unordered samples. There are n! ordered samples for each unordered sample.

Simple random sampling with replacement, however yields Nn equally likely unordered samples of which only N!/(N-n)! are of size n and the remainder have smaller sample sizes. However, when n is small compared to N, the average sample size will be smaller than n.

The reason for the common practice of ignoring repeats is that when a population unit appears in sample more than once it supplies no additional information about the population than it does when it appears once only. Orderings are similarly ignored on the grounds that the order in which a sample is selected supplies no information about the population in addition to that supplied by the values of the sample variables. The numbers of distinct population units in sample is commonly called the *sample size*, so the samples AB, AC and BC are called samples of size two units while the samples AA, BB and CC (or A, B and C) are called samples of size one unit.

There are two good reasons for preferring this sample selection procedure, which always give samples of the maximum possible size. The first is the obvious one that large samples contain more information about the population than small samples, and this affects the precision of the sample estimates. The other relates to the allocation of resources; it is nice to know before sample selection exactly how many sample units need to be collected from! Against this, however, samples selected with replacement have certain advantages in simplicity which are sometimes important enough to outweigh the two objectives just described. Because this simplicity can be quite useful at times, the current practice of ignoring orderings and repeats will not always be followed in this book. This decision will allow sampling with replacement to be examined alongside sampling without replacement and enable the student to see exactly what their advantages and disadvantages are

Two important quantities, which characterize any sample selection procedure, are the first and second order probabilities of inclusion of the population units. The first order inclusion probability of a population unit is the ordinary probability of its inclusion in sample. The second order inclusion probability of a pair of units is the joint probability of their inclusion together in sample. The first order inclusion probabilities determine the expectations (weighted averages) of estimates based on the sample; the second order probabilities determine the reliability of these estimates.

For simple random sampling without replacement (srswor) these probabilities are clearly defined and fairly readily calculable. The *first order inclusion probability* of each population unit is the same. Its probability of (inclusion) at the first draw is 1/N, at the second draw 1/N, and so on at the n-th draw it is 1/N. But it cannot be selected more than once, so its total inclusion probability is the sum of these, which is n/N.

For simple random sampling without replacement (srswor), the joint probability that a given population unit, say ith unit, is selected at the first draw is 1/N, and the conditional probability that any other population unit say jth unit is selected at the second draw given that the ith unit was selected at first denoting P(j|i) which is clearly 1/(N-1). So the joint probability that the ith unit is selected at the first draw and jth unit at the second is 1/N(N-1). By the same type of argument that was used to show that the probability of selection of any unit at any draw is 1/N, it can also be shown that the joint probability of selection of any two distinct population units at any two distinct draws in a given order is also 1/N(N-1). There are two possible orderings (ith first then jth, or the reverse) so the joint probability of selection of any two distinct draws in either order is 2/N(N-1). But there are n(n-1)/2 pairs of distinct draws (ignoring orderings) so the second order inclusion probability of any pair of distinct population units is



When we consider the concept of first and second order inclusion probabilities for simple random sampling with replacement, we have to make a choice whether to ignore repeats or not. If we choose to ignore repeats, a population unit’s first order inclusion probability is the probability of selecting it in sample at least once, which is one minus the probability of never selecting it at all, that is . When n is small compared to N, this expression is approximately equal to the simple random sampling without replacement value n/N, but it is in fact always somewhat smaller than n/N. The second order inclusion probability of any pair of distinct population units, ith and jth, is one minus the probability of not selecting ith, minus the probability of not selecting jth plus the probability of not selecting either ith or jth, that is



When n is small compared to N, this is roughly n2/N2, but once again it is always smaller than the simple random sampling without replacement value .

If, however, we choose not to ignore repeats it is necessary to move away from the strict notion of inclusion probabilities and consider instead the expected number of times a population unit appears in a sample and the expected number of times a pair of population units appear in a sample. When repeats are impossible, as for simple random sampling without replacement, the expected number of times a population unit appears in sample is its first order inclusion probability, and the expected number of times a pair of distinct population units appears in sample is that pair’s second order inclusion probability, so this new concept is really a generalization of the old one.

The expected number of times a given population unit appears in sample under srswr is







 

which is exactly the same as the simple random sampling without replacement value.

The number of times a given pair of distinct population units appears in a sample under simple random sampling with replacement is defined as the product of the times each such pair of units appears in sample in any ordering. A sample of n units (not necessarily all distinct population units) will have n(n-1) such appearances. These will be of two types. The most common type will be that in which the population units appearing at two sample draws are different, but occasionally they will be the same for any given population unit. The expected number of sample unit pairs in which that population unit appears twice will be:

There are N such population units, so an expected number of pairs of sample units is n(n-1)/N2 and n(n-1) units are paired in which the same population unit appears twice, leaving n(n-1)(N-1)/N in which different population units appear. There are N(N-1) pairs of different population units (regarding {i, j} as different from {j, i}, so the expected number of times any such pair appears is n(n-1)/N2.

Note that although the expected number of appearances of {i, j} is the same as the expected number of appearances of ith twice – both being n(n-1)/N2 – the number of times i and j appear in either order is 2n(n-1)/N2.

The example used earlier where two units were selected out of a population of three can be used to illustrate the procedure. Sampling is with replacement, so the nine possible samples are

AA AB AC BA BB BC CA CB CC.

In this array, the first letter in each pair represents the population unit selected at the first draw and the second letter the one selected at the second draw.

The pairs of sample units are now considered in both possible orderings. The sample AA therefore gives rise to n(n-1) or two identical pairs, AA and AA, but the sample AB gives rise to two different pairs AB and BA. The eighteen pairs from the nine samples are

AA, AA, AB, BA, AC, CA, BA, AB, BB, BB,

BC, CB, CA, AC, CB, BC, CC, CC.

Each of the possible pairs appears twice [that is n(n-1) times] in nine (that is N2) samples. So the expected number of appearances per sample is 2/9 [that is n(n-1)/N2]. But if the population units are distinct and the order of appearance is ignored (so that the pair AB is regarded as equivalent to the pair BA) the expected number of appearances in sample is 4/9 (that is 2n(n-1)/N2.

**2.2 NOTATION**

The following additional notation will be used in the reminder of this chapter. Note that capital letters are used for population values and lower case letters for sample values unless otherwise specified.

 The value of the variable of ith population unit

 The value of the variable of the ith sample unit

 The population mean, 

 The sample mean,  used as an estimator   
 of the population mean, 

 The population total, 

 An estimator, of the population total, 

E An operator denoting expectation

  An operator denoting variance

 An operator denoting an estimator of variance

SE An operator denoting standard error

CL An operator denoting confidence limits

 

 

**2.3 EXPECTATION OF SAMPLE MEAN**

The expectation of a random variable was defined in Chapter 1 as the sum of the products of the probabilities of a set of exhaustive and mutually exclusive events. In survey sampling, the relevant events are the selections of population units to be in the sample and the relevant values taken are the estimand (or estimated variable) values associated with the selected units. In simple random sampling both with and without replacement, the selection probabilities are all equal and the expectation is an arithmetic mean.

**THEOREM (2.2)**

In simple random sampling with replacement and without replacement, the sample mean, , is an unbiased estimator of the population mean .**◊**

**PROOF**

We select a random sample of n units from a population of N units and observe variable yi on the ith sample unit. An obvious estimator of the population mean  is

. (2.3.1)

An estimator of population total  is

. (2.3.2)

The estimator of  is unbiased estimate of  when n is large. That is to say, if we select a very large number of samples, the average value of s over all those samples would tend to . We express this formally as:



The expectation of , by definition is,

 = 

But, from Theorem (2.1), for both srswor and srswr,



where  (each unit has an equal probability to be included in the sample). Hence .**◊**

**ALTERNATIVE PROOF**

Since there are  possible distinct samples for without replacement, then

 (2.3.3)

Now in  possible samples, each unit is appearing  times, hence

*.***◊**

Similarly expected value of estimated total may be proved as an unbiased estimator of population total i.e. .

The *Bias* of an estimator is defined as the difference between the expectation of the estimator and the true value of the quantity estimated. Hence if  is used as an estimator of and the bias is zero, then  is described as *Unbiased Estimator* of. Unbiased ness is only one criterion of an estimator’s suitability, and not the most compelling one. It is easy to construct unbiased estimators, but use sometimes unsuitable for the purpose of estimation. Suppose, for instance, that an artificial variable is defined on the Ui for which the sum is zero. One such artificial variable, which is denoted by Ai, takes the value 1 for i = 1, 2,… N-1 and the value – (N-1) for i = N. Clearly the population total of the Ai, denoted by A, is zero. If  is the sample mean of the ai,formed by analogy with  then  is non – zero for all possible samples of n < N. Nevertheless  is an unbiased estimator of the population mean , which is zero. Hence all estimators of the form where c is an arbitrary constant are unbiased estimators of . If c is now allowed to become indefinitely large,  becomes inappropriate for , despite its being unbiased. To understand in what other ways  can be considered a suitable estimator for , it is necessary to consider its variance (or, since its bias is non-zero, equivalently to consider its mean squared error).

* 1. **VARIANCE OF THE SAMPLE MEAN AND ESTIMATED TOTAL**

**THEOREM (2.3)**

Suppose a simple random sample of n units is drawn from a finite population of N units. The variance of the sample mean for simple random sampling without replacement, , is

 (2.4.1)

and for simple random sampling with replacement, , is

**◊** (2.4.2)

**PROOF**

We know that

,

then



Since the variance of the sum of the random variables is equal to the sum of the variance of random variables + the sum of the covariance, we have as:

. (2.4.3)

Now regardless of whether sampling is with or without replacement,

 

  (2.4.4)

(Note that  or  is some times described as the *Population Variance* of yi.)

Further, for simple random sampling without replacement,



as )

  (2.4.5)

Substituting (2.4.4) and (2.4.5) in (2.4.3) we get



 **◊**  (2.4.1)

For simple random sampling with replacement, however, all selections are independent, therefore  is zero. Hence (2.4.3) reduces to

 (2.4.6)

Substituting (2.4.4) in (2.4.6) we obtain (2.4.2) for simple random sampling with replacement.**◊**

The variance expression for sample mean and total for simple random sampling without replacement are also written as:-

 (2.4.7)

and  . (2.4.8)

We see how (2.4.8) is built up. If there were only one unit in sample, the variance of the sample mean (that is, of the single sample observation) would be the population variance, near enough . Increasing the sample size from one unit to n units brings the variance of the sample mean down by a factor n. The estimated population total is N times the sample mean, hence the factor N2 in the formula for the variance. Finally, the fact that each unit in sample is constrained to be a different population unit means that when n = N the estimated population total must be precisely the true population total (we are ignoring all errors other than sampling error) and this has zero variance. This explains the presence of the *finite population correction* .

An analogous expression for the covariance of the unbiased estimators  and  in case of simple random sampling without replacement may be written in a straightforward way

 (2.4.9)

We see how (2.4.1) and (2.4.2) are related. Consider first the case when the sample consists of a single unit. There is then no distinction between simple random sampling with replacement and simple random sampling without replacement and the variance of  is the population variance (2.4.4). It is also the special case of (2.4.1) and (2.4.2).

When  the variance of wor is smaller than the variance of wr by a factor. Equivalently, we note that for simple random sampling without replacement the expression S2/n in the variance of  is multiplied by (1-f), which decreases as n increases, whereas for simple random sampling with replacement the same expression is always multiplied by  which does not decrease as n increases.

Another way of looking at (2.4.2) is that  in to say that the simple random sampling with replacement is same as in simple random sampling without replacement value.

People with a limited knowledge of survey sampling are sometimes under the impression that the accuracy of a sample estimator is primarily a function of the sample fraction rather than of the sample size. They would have more respect for results obtained from a sample of 5,000 out of 50,000 than they would if the results come from a sample of 20,000 out of 1,000,000, because the former is a 10 percent sample and the latter is only a 2 percent sample. Applying (2.4.1), however, we easily see that the absolute size of the sample is almost always important criterion, except when the fpc is quite large – say, greater than 0.05.

The most important point to note about the fpc is that of  is less than 2%. It reduces the variance of  by less than 20% and its standard error by less than 10% unless the fraction of population in sample is quite large, the accuracy of  entirely depend on the absolute size of the sample, n, and hardly at all on the sampling fraction. Hence, if we are comparing two populations, one containing 1,000,000 units and the other 100,000,000 units, so that any reasonably sized sample would be only a small fraction of either. We would naturally think of choosing the same sized sample from each population, and do not worry that we have selected a much thinner sample from the large population than from the smaller one.

**2.4.1 Alternative Proof Using Indicator Variable**

The following alternative proof uses *Indicator Variables*. The use of indicator variable is particularly convenient in unequal probability sampling (ups) or probability proportional to size sampling (pps). Readers who intend to master the topic of pps sampling are advised to familiarize themselves with its use in simple random sampling first.

Indicator variables are random variables, which indicate the number of times a population unit is selected in the sample. Formally the indicator variable is denoted by ai, where

ai  = ni if ith unit of the population is in sample ni times.

In simple random sampling without replacement ni takes the value 1 with probability f=n/N and 0 with probability . Further, the joint probability that ai = 1 and aj = 1 is n(n-1)/N(N-1), for simple random sampling without replacement.

Now

 (2.4.10)

 (2.4.11)

and



  (2.4.12)

In simple random sampling with replacement, however, ai(s) have the multinomial distribution that arises wherever there are n independent trials having N equi-probable outcomes. Here we have

 (2.4.13)

 (2.4.14)

and

 (2.4.15)

For both simple random sampling without replacement and simple random sampling with replacement the sample means can be written using indicator variable notation as:

 (2.4.16)

Note that in (2.4.16) the ai(s) are random variables and the Yi (s) are fixed values. The expectation and variance of may therefore be written

 (2.4.17)

and

 (2.4.18)

For simple random sampling without replacement substitute the value of  from (2.4.10) in (2.4.17)

, (2.3.1)

which proves the unbiased-ness of .

Substituting the value of  and  from (2.4.11) and (2.4.12) in (2.4.18) we get



 **◊**  (2.4.1)

For simple random sampling with replacement, substituting (2.4.13) in (2.4.17), we get



Again substituting (2.4.14) and (2.4.15) in (2.4.16) we have for simple random sampling with replacement



 **◊** (2.4.2)

The variance of the estimated total  for both simple random sampling with and without replacement is

 (2.4.19)

**2.5 ESTIMATOR OF THE VARIANCE OF THE SAMPLE MEAN**

The variance expressions (2.4.1) and (2.4.2) contain the expression S2, which is a function of all the population values Yi. Since only those values, which relate to the units in the sample, are known, We need to estimate variances using only those values that fall in the sample. We therefore have to have an estimator of the variance of the estimator of population mean.

**THEOREM (2.4)**

Suppose a simple random sample of n units is drawn from a finite population of N units. In unbiased estimator of the variance of the sample mean, , is

 (2.5.1)

if sampling is without replacement and

 (2.5.2)

if sampling is with replacement.**◊**

The proof of this theorem requires the following lemma:

**LEMMA (2.4.1)**

For simple random sampling without replacement, s2 is an unbiased estimator of S2 and for simple random sampling with replacement s2 is an unbiased estimator of S2 (N-1)/N.**◊**

# PROOF

For both simple random sampling without replacement and simple random sampling with replacement, we have

 







. (2.5.3)

Taking the expectation of (2.5.3) we have

.

Since from (2.4.4) we know that therefore 

Now  is the variance of the sample mean, which is  for simple random sampling without replacement and is  for simple random sampling with replacement. Hence for simple random sampling without replacement

 (2.5.4)

and for simple random sampling with replacement

 **◊** (2.5.5)

The Theorem 2.4 can be proved as:

**2.5.1 Proof of Theorem 2.4**

For simple random sampling without replacement, the expectation of (2.5.1) is

**◊**

For simple random sampling with replacement, the expectation of (2.5.2) is

**◊**

**2.5.2 Alternate Expression of Variance for Sample mean**

s2 may be written alternatively as:

, (2.5.6)

and

. (2.5.7)

Consequently, the basic building block, which go to make up s2 are therefore the comparisons between all pairs of sample units. The same is true of S2.

 (2.5.8)

or

. (2.5.9)

This will be of considerable importance when we consider systematic sampling.

**2.6 STANDARD ERRORS AND CONFIDENCE LIMITS**

**FOR THE POPULATION MEAN**

**2.6.1 Standard Error.**

Standard error, in fact , estimate the amount of sampling error which comes in the estimation process because of sampling process. The standard error of the sample mean is the square root of the sampling distribution of sample mean i.e., consequently the standard error of the sample mean  is



 (2.6.1)

**2.6.2 Confidence Limits of Sample Mean**

It is not possible for a sample to evaluate characteristics of a population exactly, but it estimates the characteristics as accurately as possible. One way out may be to find an interval, which covers the parameter with some pre-assigned probability. In case the sample estimate is the arithmetic mean, and observations are normally distributed with known variance, the sampling distribution of sample mean,  is normal with mean  and variance given by in the expression (2.4.1) or (2.4.2). If samples are repeated, then the interval covers  in 95% of the cases. This can be stated as



which shows that the probability of  lying between



is 95%. This interval is called a *confidence interval* and the probability attached to it is known as *confidence coefficient*.

In general for any probability level (1- α) the confidence limits for sample meanwith unknown variance is

 (2.6.2)

and the confidence limits for estimated population total 

, (2.6.3)

where t is the t-probability point corresponding to the desired probability. Under fairly general conditions, for non-normal populations the sampling distribution of sample mean is approximately normal provided the sample size is sufficiently large and as such the procedure for setting up the confidence limits is the same as in case of normal population. Upper and lower confidence limits are random variables. The probability that it lies above its upper (1 – α) percent confidence limit is only α percent. Similarly the probability that it lies below its lower (1 – α) percent confidence limit is also α percent. Conventionally, the most used values of α are 0.01 and 0.05. If α = 0.05 we have 95 percent confidence limits and if α = 0.01 we have 99 percent confidence limits. On a repeated sampling basis we have 19 chances in 20 that a random variable will be observed to lie between its 95 percent confidence limits, and 99 chances in 100 that it will be observed to lie between its 99 percent confidence limits.

In order to estimate confidence interval for a population mean in terms of standard error or estimated standard error, it is necessary to make some assumptions about the sample mean. The simplest and most usual assumption is that this distribution is approximately normal. Whether it in fact is approximately normal depends both on the distribution of the Yi values in the original population and on the sample size n. since the variance of the Yi is necessarily finite, the central limit theorem ensures that under simple random sampling with replacement the distribution of  is asymptotically normal. The non-normality of this distribution (as measured by the ‘excess of kurtosis’ over the normal value of 3) diminishes as 1/n. But if the original population differs a great deal from normality, this may not be necessarily fast enough for practical purposes.

**Example 2.1:**

Draw all possible samples of size 2, with and without replacement from a hypothetical population of 4 units. Verify that;  for without replacement and  for with replacement. Also find the variance of  and compare the result with that obtained by formulas (2.4.1) and (2.4.2). The values of the population are: Yi= 2, 4, 6, 8.

## SOLUTION

Population mean  and S2 = 6.67

i) Possible samples for with replacement .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 2 | 4 | 6 | 8 |
| 2 | (2, 2) | (2, 4) | (2, 6) | (2, 8) |
| 4 | (4, 2) | (4, 4) | (4, 6) | (4, 8) |
| 6 | (6, 2) | (6, 4) | (6, 6) | (6, 8) |
| 8 | (8, 2) | (8, 4) | (8, 6) | (8, 8) |

The  and  for all possible samples as

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|  | 2 | 3 | 4 | 5 | 3 | 4 | 5 | 6 | 4 | 5 | 6 | 7 | 5 | 6 | 7 | 8 |
|  | 0 | 2.0 | 8.0 | 18.0 | 2.0 | 0 | 2.0 | 8.0 | 8.0 | 2.0 | 0 | 2.0 | 18.0 | 8.0 | 2.0 | 0 |
| P | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 |

 and 





ii) For without replacement

Since N = 4; n = 2, the possible samples under without replacement sampling scheme is  All possible samples, mean and s2 are givenas::

|  |  |  |  |
| --- | --- | --- | --- |
| Possible samples |  | **s2** | P(s) |
| 2,4 | 3.0 | 2.0 | 1/6 |
| 2,6 | 4.0 | 8.0 | 1/6 |
| 2,8 | 5.0 | 18.0 | 1/6 |
| 4,6 | 5.0 | 2.0 | 1/6 |
| 4,8 | 6.0 | 8.0 | 1/6 |
| 6,8 | 7.0 | 2.0 | 1/6 |
|  | 30.0 | 40.0 |  |





If we use expressions (2.4.1) and (2.4.2) we get 

,

and  for without replacement and with replacement respectively. The standard errors of sample mean for without replacement and with replacement respectively are:

 and .

**Example 2.2:**

A simple random sample of 33 mohallahs was taken from a population of 663 mohallahs (from the first report of 1973 population census of Libya). The population (male and female) of the sampled mohallahs is given as:

3485, 3085, 2049, 2952, 1269, 754, 449, 1722, 6907, 984, 3890, 2521, 1394, 4737, 4956, 3671, 766, 1505, 14901, 1111, 1362, 2136, 2200, 14123, 1347, 458, 1923, 2793, 18585, 22422, 842, 2575, 6341.

1. Estimate average population per mohallah, [the actual average population of 663 mohallah is 3404 persons].
2. Estimate the total population of Libya; [the actual population of Libya for 663 mohallahs is 2,257,088].
3. Find the  [the s2 of the population is 25,242,133.05]
4. Find confidence limits for sample mean and total.

## SOLUTION

For this example N = 663, n = 33, Σyi = 140251, (Σyi)2/n = 59,576,530.3



1. The average population per mohallah is 140251/33 = 4250.03.
2. Estimated total of 663 mohallahs 
3. 
4. Standard error of the mean = 934.818
5. Standard error of total is 
6. 



vi)  



Further three samples (5%, 10% and 20%) were randomly selected from 663 mohallahs of Libya. Means, estimated total and their standard errors are computed and are given as:

|  |  |  |  |
| --- | --- | --- | --- |
| Sample % | **5%** | **10%** | **20%** |
| Sample units | 33 | 66 | 132 |
| Average population per mohallah | 4250 | 3969 | 3331 |
| Estimated total | 2817750 | 2631512 | 2209000 |
|  | 934.82 | 688 | 374 |
|  | 619784.33 | 456064 | 247,735 |

One can say that as the size of the sample increases standard error of sample mean decreases. Consequently one can say that as the sample size increases, sampling error decreases.

**Example 2.3**

A population of Multan district of males and females of 523 villages is given at the end of this chapter. The total, mean and variance of Male, Females and Total population are calculated and given as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sex | N0.villages |  | Mean | s2 |
| Males | 523 | 937784 | 1793.29 | 3867810.5 |
| Females | 523 | 864058 | 1793.09 | 3281968.7 |
| Total | 523 | 1800841 | 3443.29 | 14253440.5 |

The samples of 5%, 7% and 10 % have been selected using SPSS Package (random number samples may also be drawn by using random tables given at the end of this chapter). The villages selected on the basis of 5% are given below. But the results of the samples for 7 % and 10% are given only. Estimate the total population on the basis of these results and calculate the 95% confidence limits for population mean

**Solution**

The 5% sample using the random process has been selected and population for males, females and total of males and females is given below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Males** | **Females** | **Total** | **Males** | **Females** | **Total** |
| 260 | 228 | 488 | 718 | 658 | 1376 |
| 180 | 152 | 332 | 870 | 843 | 1713 |
| 1542 | 1348 | 2890 | 432 | 401 | 833 |
| 2987 | 2926 | 5913 | 281 | 228 | 509 |
| 3231 | 2934 | 6165 | 662 | 587 | 1249 |
| 994 | 857 | 1851 | 700 | 699 | 1399 |
| 18913 | 17155 | 36068 | 1850 | 1758 | 3608 |
| 1303 | 1182 | 2485 | 266 | 217 | 483 |
| 177 | 171 | 348 | 416 | 365 | 781 |
| 1299 | 1224 | 2523 | 692 | 631 | 1323 |
| 8625 | 8072 | 16697 | 3061 | 2883 | 5944 |
| 3164 | 3072 | 6236 | 668 | 660 | 1328 |
| 1502 | 1466 | 2968 | 300 | 300 | 276 |

The mean, variance and standard error for males, females and total are calculated and given as

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sex |  |  | SE of () | s2 |
| Males | 55093 | 2118.96 | 754.86 | 14815215.5 |
| Females | 51017 | 1962.19 | 688.76 | 14333971.7 |
| total | 105786 | 4068.69 | 1444.79 | 54272633.8 |

95% confidence limits for males, females and total are as:

C.L ( = males) = 639.44 ~ 3598.48

C.L ( = females) = 612.22 ~ 3312.16

C.L ( = total) = 1179.11 ~ 6958.27

A sample of 7 % was taken from a population of 523 villages and the statistic calculated is as

|  |  |  |  |
| --- | --- | --- | --- |
| Sex |  |  | Standard Error of |
| Males | 60112 | 1624.65 | 321.81 |
| Females | 55302 | 1494.65 | 297.87 |
| Total | 115414 | 3119.30 | 619.60 |

95% confidence limits for males, females and total are as:

C.L ( = males) = 944.65 ~ 2255.4

C.L ( = females) = 910.82 ~ 2077.8

C.L ( = total) = 1904.88 ~ 4333.72

Another sample of 10% was selected from this population relevant information s:

|  |  |  |  |
| --- | --- | --- | --- |
| Sex | Total | Mean | Standard Error of |
| Males | 94615 | 1819.52 | 239.06 |
| Females | 86643 | 1666.21 | 222.86 |
| Total | 181258 | 3485.73 | 461.48 |

C.L ( = males) = 1332.96 ~ 2288.08

C.L ( = females) = 1229.71 ~ 2103.02

C.L ( = total) = 2581.23 ~ 4389.73

The estimated total its standard error and confidence limits are given as:

|  |  |  |  |
| --- | --- | --- | --- |
| Sample size % |  | SE of () | C.L of () |
| 5 | 2127925 | 755625.17 | 646980 - 3608950 |
| 7 | 1639394 | 324050.8 | 1004254 - 2274534 |
| 10 | 1823037 | 241354.04 | 1349983 - 2296090 |

Some important remarks regarding their calculations are:

1. All the confidence limits contain respective population mean.
2. As the sample sizes increases standard error decreases.

**Example 2.4:**

The sample mean,  is the best linear unbiased estimator of population mean based on a simple random sampling of size n.

**SOLUTION:**

Let us consider any linear unbiased estimator



The variance of t can be written as



Substituting the value of  and from (2.4.4) and (2.4.5) we get

 





Since,  therefore



The problem is to minimize  subject to . By Lagrange multiplier (λ ) we minimize  unconditionally



Differentiation the above w.r.t. ai we get

= Constant = c or .

Since , therefore nc = 1 or c = 1/n or ai = 1/n

In other words ai is constant. The unbiased ness condition of E(t) =  shows that  Hence  and 

We say that the sample mean is the best linear unbiased estimator of population mean.

**2.7. ESTIMATION FOR PROPORTIONS**

Sometimes results are required in the form of proportions and percentages. In this section the estimation of proportion (say P) of units of population belonging to a special class is considered. In medical, biological and social sciences results are needed in the form of proportion and percentage. P may be defined as a proportion or a percentage in the population, e.g percentage of unemployed people in a certain country, proportion of villages having medical facilities, proportion of educated people or percentage of people suffering from a particular disease or proportion of smoker in some class. P may be defined as proportion of persons suffering from a certain disease etc.

Suppose we have N population units i.e. Y1, Y2, …. Yi, …YN. yi = 1 if the unit possesses a certain attribute and 0 otherwise. If some numbers (say A) are falling in a particular class, then the proportion of these units may be defined as P = A/N. Similarly if we have n sample units i.e. y1, y2, …. yi, …. yn, yi = 1 if the unit possesses a certain attribute 0 otherwise, then proportion of units possessing the attribute may be defined as p = a/n.

Using the above definitions, the value of Yi is either 1 or 0, then

 (2.7.1)

# Since Yi takes the value 1 or 0; also takes the value 1 or 0. Thus

 (2.7.2)

The value of S2 and s2 in terms of proportions may also be derived as:

We know 

Using (2.7.1) and (2.7.2), we have

 (2.7.3)

Similarly s2 = npq/(n – 1) (2.7.4)

**THEOREM (2.5)**

T

he sample proportions, p, is an unbiased estimator of population proportion P both for simple random sampling without replacement and simple random sampling with replacement.**◊**

The Proof is trivial.

**2.8. VARIANCE OF SAMPLE PROPORTION AND ESTIMATED TOTAL**

**THEOREM (2.6)**

A simple random sample of size n is drawn from a finite population of size N. The variance of the proportion p for simple random sampling without replacement is

 (2.8.1)

and for sampling with replacement is

**◊** (2.8.2)

**PROOF**

We know that

 (2.4.1)

Both for srswor and srsws we use (2.7.3) in (2.4.1),

**,**

and

Var (pwr) = PQ/n **◊**

The unbiased variance estimators of p may also be obtained by using (2.7.4) on the similar lines:

 (2.8.3)

 (2.8.4)

If fpc is ignored and n is large then (2.8.3) and (2.8.4) are equal to

. (2.8.5)

It matters little that for large sample size n, whether to use (n – 1) or n in the denominator of expression (2.8.5). Hence we may use

. (2.8.6)

The estimated total A′ and Var(A′) may respectively be obtained as

A′ = Np and (2.8.7)

Var(A′) = N2 Var(p). (2.8.8)

**2.8.2. Standard Error and Confidence Limits of p**

The standard error of p and A′ (estimated total = NP) may be obtained as

 (2.8.9)

and

 (2.8.10)

and confidence limits for p and A′ are

C.L(p) = p +  S.E(p) (2.8.11)

and

C.L(A′) = A′ +  S.E(A′). (2.8.12)

Also in presenting the results, the units are classified with more than two classes. Then

, , , .

If is constant, then the probability of drawing the sample is given by the multinomial expression

, (2.8.13)

which is an extension of the binomial distribution.

## EXAMPLE 2.5

In a simple random sample of 200 colleges, 120 colleges are in favor of a proposal, 57 opposed and 23 do not express their opinion. Estimate the total number of colleges in the population who are in favor of the proposal and find 95% confidence limits for the number of colleges, which favore the proposal. The total number of colleges in the population is 2000. (Source: Cochran 1977)

### SOLUTION

We have N = 2000 and n = 200. Since 120 colleges are in favour of a proposal, hence p = 120/200 = 0.6 and q = 0.4. The estimation of colleges which are in favour of the proposal from 2000 will be Np = 0.6 x 2000 = 1200.

The standard error of the estimated total is



The confidence limits are

A′ + t S.E(A′) = 1200 + 1.96 x 69.2= { 1064, 1335 }.

One can say that these limits contain true population proportion at 95% probability.

### EXAMPLE 2.6

A survey is conducted in the Faculty of Science, El-Fateh University to estimate the proportion of smokers. A simple random sample of 122 students is taken from a population of 813 students. The total number of smokers in the sample was 42. Estimate the total number of smokers in the Faculty of Science, with 95% confidence limits. Among 122 students 86 were males and 36 were females and there was no female smoker. Find the proportion of smokers ignoring the female students. Also construct 95% confidence limits of this estimate. The total number of males in the faculty was 551.

### SOLUTION

Since there are 813 students and a sample of 122 is taken from this population, so   
N = 813, n = 122.

Out of 122 only 42 are smokers, hence the proportion of smokers is

p = 42/122 = 0.3443, q = 0.6557

Since the sample is large f.p.c. may be ignored. The standard error of p is:



The estimated total number of smokers is

Np = 813 x 0.3443 = 279.916 = 280.

The standard error of the estimated (A′) total is

S.E(A′) = 813 x 0.043 = 34.96

95% confidence limits for the total are

280 + 1.96 x 34.96 = (211 ~ 349).

Now there is no female smoker, the proportion of smoker is

p = 42/86 = 0.4883, q = 0.5117 S.E(p) = 0.0539. The standard error of the total is

S.E(A′) = .0539 x 551 = 26.699.

The confidence limits for proportion are

0.4883 × 551 + 1.96 × 29.699 = (211 ~ 327).

**2.9 ESTIMATION OF SAMPLE SIZE.**

In a sample survey, it is always a problem for an experimenter to know or to estimate the size of the sample when the result is required with least sampling error. It is always a problem whether a sample should be 2%, 5%, or 10% or any other fraction. Although the sample size has always been a matter of choice with planners, yet great care and weight is needed in its estimation. Since sample is a proportion of the population, it should neither be too large to involve a lot of expenditure nor too small to make the result less reliable. In fact the sample size depends on the cost involved and precision required.

Let ‘d’ be the margin of error with some probability α by which sampling value differs from population value. The permissible error, which is a percentage difference between the estimate and parameter value, is specified as:

 (2.9.1)

Since differs from sample to sample, the probability of the margin of error being less than d is given by

 (2.9.2)

α is usually chosen such that 1 - α is 90% or 95% or some other desired level. Let us assume that  is normally distributed with mean  then

 (2.9.3)

so that

. (2.9.4)

Now consider the case of sampling without replacement. Substituting  in (2.9.4) we have

. (2.9.5)

Solve for n, we have  (2.9.6)

For large N

, (2.9.7)

Then . (2.9.8)

If N is large (2.9.6) is identical to (2.9.8). Denoting n0 = (tS/d)2, as the first approximation, the second approximation is

 (2.9.9)

the third approximation may be found as

 (2.9.10)

and so on.

**Example 2.6**: A physician would like to know the mean fasting blood glucose of patients seen in the diabetes clinic over the past 10 years. Determine the number of records the physician should examine in order to obtain 90% (and 95%) confidence level for population if the desired width of the interval is 8 units. A Pilot sample yields a standard deviation of 60 units.

**Solution:**

Here s = 60, d = 4, as the total width is 8 which is on the both sides of the mean. Therefore, the sample size for 90% confidence is, n =  = 609, and for 95 %, n =  = 864.

**EXAMPLE 2.7**

According to population census of Multan district (Appendix II) the variation of population males and females is 3867810.5. Suppose we like to hold a census on sample basis to estimate the population of Multan Dist. What would be the sample size, when the total number of villages in this district is 523. Use 95% probability confidence coefficient.

**SOLUTION:**

S2 = 3867810.5, N = 523.

The range of distribution of population is very large and a lot of variation is in the population of the village therefore we can take d = 0.2 (20% approximately). Using (2.9.6) we can start the approximation and then by iterative method we proceed as:

First approximation: = 523

Second approximation: 

Third approximation: 

This process is repeated and







Since the difference between last three approximations is constant so 44 is the optimal sample size.

**2.9.1 Sample Size for Coefficient of Variation**

In case of coefficient of variation, the sample size may be determined as follows:

For large N the coefficient of variation of sample mean is



or (2.9.11)

or (2.9.12)

Since  is coefficient of variance per unit, therefore,

 (2.9.13)

Note: if a sample of size n gives C.V = C1for a sample estimate, then the sample estimator with C.V. = C2, the required sample size is,

n1 = n(C1/C2)2 (2.9.14)

Comparing (2.9.13) and (2.9.14) we get

, (2.9.15)

where n is the size of the first sample.

##### EXAMPLE 2.8

Suppose it is believed that about 20% of persons in a large population suffer from a certain disease. How many persons should be selected in a random sample in order that the coefficient of variation of the estimate be 10% or less.

##### SOLUTION

We know that

 C.V = 1/10, since p = 0.2 and q = 0.8, 

hence the sample size is 400.

##### EXAMPLE 2.9

A simple random sample of 50 households is taken from a population of a certain village for the estimation of expenditure on meat. The estimated average expenditure on meat per household comes out Rs. 0.88 with standard error of Rs. 0.10 Rs. Using this information find out what sample size is needed if similar type of survey is to be carried out in some other village with permissible 95% probability level is 10% of the true value.

##### SOLUTION

The permissible margin of error at 95% probability is 10% of the true value,

hence

0.1 0  = t S.E

S.E = 1.96 S.E(

S.E = 0.051 = C2

We assume that the C.V. is the same in both the villages.

The C.V. of the first village is 0.1/0.88 = 0.1136 = C1

Hence

n1 = n(C1­/C2)2 = 50(0.1136/0.051)2 = 248

**2.9.2 Sample Size for proportion**

The determination of sample size in case of proportion may be found to as.

d = t S.E(p) (2.9.16)

Using (2.7.3) (2.7.4) in (2.9.16) we obtain

 (2.9.17)

For large N

. (2.9.18)

Solving (2.9.17) and (2.9.18) for n we get,

 (2.9.19)

and

n = t2PQ/d2. (2.9.20)

respectively.

For practical purposes population values are replaced by sample values i.e. s for S and p for P at the respective places. Denoting n0 = t2PQ/d2 the second approximation from (2.9.19) may be

 (2.9.21)

and so on.

**2.9.2.1 *Sample size for absolute precision***

**Example 2.10**:

The Ministry of Health wishes to estimate the prevalence of tuberculosis among children under 5 years of age. How many children should be taken in the sample so that the prevalence may be estimated within 5% points of the true value with 95% confidence level, if it is known that the true rate does not exceed 15%.

**Solution:**

In this example we have

p = 0.15, 1–p = 0.85

Probability level or confidence level (1 – α) = 95%.

d = 5 percentage points

 and 

Using the formula, we have

 = 196 for 1– α = 95%

If population is finite then an approximation of sample size can be obtained as   
n1 = . If the population of children less than 5 years of age is 20,000, then the sample size may be estimated as, by an iteration,

 =  = 194

This is not different from 196, so 196 or 194 may be taken as a sample size.

**Example 2.11**

Ministry of Health would like to estimate the proportion of children who are receiving medical care regularly. How large should be the sample if the estimate falls within 5% of true proportion with 95% confidence level.

**Solution:**

In this question, the assumption regarding proportion of children who are receiving regularly medical care is that 50% of the population of children is receiving medical care. Using p = 0.50, maximum sample size will be obtained.

If we take

p = 0.5; 1 - α = 0.95, 0.99; d = 0.05 ,then

 = 384 for 95%

 = 666 for 99%

Suppose N = 600 than, then the sample size for 95% level comes to be:

n1 =  =  = 234 (2nd approx.)

n2 =  =  = 169 (3rd approx.)

n3 =  =  = 132 (4th approx.)

This process will continue till difference between the last two approximations becomes minimal.

**2.9.2.2 *Sample size for relative precision***

If the coefficient of variation (or for relative precision) is given, the formula for the determination of sample size is

n = , **(**2.9.22)

where D denotes coefficient of variation or relative precision.

For convenience, sample sizes have been calculated for different values of p and D. [see Tables at the end of this chapter]

**Example 2.12**

Ministry of Health of Eastern Province would like to conduct a survey regarding hypertension of elderly persons (above the age of 60). It is known from the past experience that the prevalence of hypertension is 25%. How large a sample should be so that the resulting estimates falls within 10% (not 10% points) of the true proportion with 95% confidence level?

**Solution:**

In this question p = 0.25, Confidence level = 95% and relative precision is 10% of 25%. There are two ways to solve this problem.

(i) Using relative precision formula

n =  = 4610

(ii) Using absolute precision formula

Since d = 0.05 x 0.25 = 0.0125

n =  = 4610

If population size is known to be 2000, then

n1 =  =  = 1395

n2 =  =  = 822

n3 =  =  = 583

This process will continue till there is not much difference between the last two approximations. We see that after 10th approximation, we get the sample size of 212.

If p = 25% to 40% and relative precision D = 0.05 then for different values of p and with 95% confidence level, the sample sizes are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| P | 0.25 | 0.30 | 0.35 | 0.40 |
| N | 4610 | 3585 | 2854 | 2305 |

The relative precision (D) may be converted into absolute precision (d) as

d = p x D = 

The sample sizes for different values of d and p and for 95% confidence level are given as:

Sample sizes for different values of p and d

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| d →  p↓ | 0.0125 | 0.0150 | 0.0175 | 0.0200 |
| 0.25 | **4610** | 3201 | 2352 | 1801 |
| 0.30 | 5163 | **3585** | 2634 | 2017 |
| 0.35 | 5593 | 3884 | **2854** | 2184 |
| 0.40 | 5901 | 4098 | 3010 | **2305** |

If the range is given, i.e. the prevalence is 10 to 25%, then it is always advisable to use prevalence 25% for precision. If the range is 45% to 55% then for absolute precision use p = 50% but for relative precision use 55%.

In case the coefficient of variation is given, the formula for the determination of sample size will be

 (2.9.22)

##### EXAMPLE 2.13

A survey is to be made of the common disease in a large population. For any disease that affects at least 1% of the individual in the population, it is desired to estimate the total number of cases, with a coefficient of variation of not more than 20% (a) what size of sample is needed, assuming that the presence of the disease can be recognized without mistake (b) what size is needed if the total cases are wanted for male and female with the same precision (Source: Cochran, 1977).

##### SOLUTION

Since

p = 0.01 and q = 0.99, C.V = 0.20

C.V(per unit) = 

Using

 = .99/.04 = 2475

If males and females are needed separately with same C.V two samples are assumed to be taken and probability of getting disease for male and females is same; hence the sample size will be 2475 for each.

It is hard to estimate the sample size without some prior knowledge about the S or C.V. This is the reason that sampling statisticians always record S or C.V. for which they worked. If S or P are not known, their values can be estimated by four possible different ways.

1. from some previous survey
2. from the experience of sampling statistician
3. from pilot studies

**2.9.1 Estimation of Sample Size When Cost is Involved**

Since cost plays a major role in the conduct of surveys, so in the estimation of sample size, cost aspect should be taken into account. Let the simple linear cost function is

C = C0 + nC1, (2.9.23)

where C = total cost, C0 = overhead cost, C1 = cost per unit.

Let the loss due to  not being equal to  is

, (2.9.24)

where d is constant.

The total survey cost and the loss involved will be

 (2.9.25)



For fixed sample size (2.9.25) is

 (2.9.26)

If L(n) is plotted against n, then the graph will have minimum value at some value of n. This value of n is termed as optimum value of sample size. The optimum value of n may be obtained by the derivative of (2.9.26) with respect to n.

After ignoring finite population correction the . Putting the value of  in (2.9.26), we have

L(n) = C0 + C1n + d S2/n. (2.9.27)

Finding the differentiation of the above expression w.r.t. n and equating to zero, the value of n that minimizes cost is

 (2.9.28)

##### EXAMPLE 2.14

If the loss function due to not being equal to  and the cost function is C = C0 + C1 n show that the sample size in case of simple random sampling ignoring f.p.c. is

.

###### SOLUTION

Since the loss is given in terms of mean deviation, the loss due to  not being equal to  will be



The total survey cost and the loss involved is

L(n) = C0 + C­1n + 

Finding the derivatives with respect to n and equating to zero, we have

.

**Example 2.15**

A sample of size n is selected from a population of N units; another sample of the same size is selected from the remaining N – n units. If and are the means of the first and the second sample and  is the population mean then

1. ,
2. ,
3. ,
4. If  is the pooled estimator, then



which is a variance of sample size of 2n taken from a population units.

## SOLUTION

1. Let  be the mean of a sample of size n taken from a population of N units then we have already proved that 
2.  mean of the remaining units





1. 

, where









Now

 

Therefore 

 

(vi) 

 as





## Alternative Proof of Example 2.15

Let the two estimators be:



where ai is an indicator variable defined earlier.

Now 



Since  is mean of sample which is selected after the selection of first sample, therefore if can be written as



Now 





Again 



Again consider 

or 

or 

 where aK will have 2n values.

Now 

 









**EXAMPLE 2.17**

To estimate the total employment in an industry comprising of 70 factories, a sample of 10 factories is selected in the following manner. Three factories with relatively larger number of employees are definitely selected for the survey and a sample of 7 factories is selected by using srswor from the remaining 67 factories. Two estimators of the total employment in that industry are proposed:

(i)  and (ii) ,

where wi denotes the number of employees in the ith sample factory. Compare the mean square error of using the following data:

|  |  |
| --- | --- |
| Total number of employees for the 3 big factories | 8050 |
| Total number of employees for the remaining 67 factories | 60810 |
| Sum of squares of the number of employees for the remaining 67 factories | 71740058 |

(Source: Murthy 1966)

## SOLUTION:

Given estimators are:

(i)  and (ii) 

Total number of employees in 3 big factories and other 67 factories are 8050 + 60810 = 68860 = Y.





 



so is a biased estimator.

Bias in   = 100823 – 68860 = 31963.



 = 8050 + 60810 = 68860 = Y.

Since, therefore  is an unbiased estimator of Y.



 

 



 = 77,015,256

 = 1098648600

Since there is no bias in 



 = 143,990,660 

.

**EXERCISES**

1. From a population of N = 6 units, draw all possible samples of size 2 and 3 and demonstrate that  also E(s2) = S2 the population values are Yi = 3, 4, 6, 7, 9, 12.
2. A simple random sample of 100 households is selected from the village having 850 households for the record of television sets. Survey record shows that  and  Estimate total T.V. sets in 850 houses. Find 95% confidence limits.
3. A simple random sample without replacement of 10 agriculture plots was taken from 100 plots of a village and areas under wheat was recorded 4.5, 2.5, 3.2, 2.6, 1.9, 2.8, 3.1, 2.8, 3.0, 2.8 acres. Estimate the total cultivated area under wheat. Find the standard error of the total mean under wheat and confidence limits.
4. A simple random sample of 2070 farms is taken from a population of 207 00 farms and information is collected on the number of cattle (yi) on each farm. The following data are obtained;  Estimate total number of cattle in the population. Find the coefficient of variation. How many farms should be more selected in a future survey in order that the relative error of the estimate be not more than 5%.

(i) 258810 (ii) 336

1. Four simple random samples of different sizes are taken from 663 mohallah from the 1973 population census of Libya. The values are recorded in the following table (yi denotes the agriculture holder). Estimate the total agricultural holders and find the standard error in each case. The true total is 165541.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample | 10% | 5% | 10% | 20% |
| Units | 7 | 33 | 66 | 132 |
| Σyi | 1080 | 8219 | 15538 | 34801 |
| Σy21 | 240592 | 3075517 | 5446659 | 15276357 |

Comments on the finding of result.

1. In a sample of 200 colleges from a population of 2000, 140 colleges were in favour of the proposal that a new system of examination should be introduced (i) estimate the 95% confidence limits for the number of colleges in the population in the favour of the proposal (ii) do the above data furnish evidence at 95% level to reject the hypothesis that the only 50% of all the colleges are in favour of this proposal.
2. A simple random sample of 290 households was chosen from a city are containing 14,828 households. Each family was asked whether it owned or rented the house and also whether it had the exclusive use of an indoor toilet. Results are as;

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Owned** | | Rented | | **Total** |
|  | **Yes** | **No** | **Yes** | **No** |
| Exclusive use of toilet | 141 | 6 | 109 | 34 | 290 |

For the family who rent, estimate the percentage in the area with exclusive use of an indoor toilet and give the standard error of this estimate. Estimate the total number of renting families in the area who do not have exclusive indoor toilet facilities and give the standard error of this estimate. If the total number of renting families in the city area is 7526, make a new estimate of the number without toilet facilities and give the standard error of this estimate.

1. Material for the construction of 5000 wells was issued in a district. The list of cultivators whom the material was issued available along with the proposed location of each well. A large part of material was, however, reported to have been misused. It is proposed to estimate the proportion of wells not actually constructed, by taking a simple random sample wells with the permissible margin of error of 10% and the degree of assurance desired 95%. Determine the size of the sample required to estimate the proportion of wells not constructed for different values of the population proportion ranging from 0.5 to 0.9.
2. A simple random sample of 50 households was selected from 300 households. Out of 50 households 10 were found to possess a T.V. sets. Members of the families were 4, 4, 5, 5, 6, 5, 6, 7, 3, and 5. Estimate the followings:
3. total number of households in the population possessing T.V. sets,
4. total number of persons in the population.
5. A population contains N units, the variate value of one unit being known to be Y­1­. A without replacement sample of size n is selected from the remaining (N-1) units. Show that the estimator  has a smaller variance that  based on a random sample without replacement of size n taken from the whole population.
6. Select a simple random sample of size 10 from the given 47 villages of a certain districts of Punjab area, showing area under guava crop. Estimate the total area under guava crop. This distinct along with standard error of your estimate compare your result with actual.

|  |  |  |  |
| --- | --- | --- | --- |
| Selected  Village | Area Under  Guava Crop | Selected  Village | Area Under  Guava Crop |
| 01 | 166.15 |  | 12.57 |
| 02 | 24.73 |  | 2.00 |
| 03 | 100.77 |  | 6.72 |
| 04 | 87.14 |  | 20.75 |
| 05 | 116.28 |  | 51.65 |
| 06 | 60.22 |  | 16.42 |
| 07 | 13.59 |  | 3.90 |
| 08 | 41.70 |  | 2.44 |
| 09 | 10.52 |  | 3.90 |
| 10 | 13.85 |  | 15.31 |
| 11 | 12.92 |  | 1.44 |
| 12 | 10.73 |  | 14.88 |
| 13 | 38.64 |  | 23.01 |
| 14 | 15.92 |  | 3.44 |
| 15 | 9.09 |  | 14.32 |
| 16 | 155.51 |  | 24.39 |
| 17 | 10.34 |  | 9.88 |
| 18 | 95.16 |  | 17.66 |
| 19 | 22.40 |  | 3.26 |
| 20 | 10.97 |  | 6.89 |
| 21 | 39.07 |  | 0.84 |
| 22 | 13.70 |  | 13.02 |
| 23 | 26.64 |  | 32.85 |
| 24 | 1.40 |  |  |

1. The data given below pertain to one complete location of milk yield of 250 cows in an organized daily farm.
2. Select a simple random of size 25.
3. Estimate the mean with its standard error.
4. Construct a 95% confidence limit for the population mean.

Milk yield (in 10 kgs.) of 250 cows of one complete location (305 days) in an organized dairy farm (units are numbered column wise). {source Sing, Sing and Kumar 1981}.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 230 | 293 | 163 | 290 | 200 | 173 | 194 | 322 | 169 | 230 |
| 297 | 151 | 248 | 271 | 269 | 214 | 167 | 207 | 240 | 286 |
| 184 | 248 | 327 | 338 | 165 | 177 | 270 | 177 | 202 | 155 |
| 155 | 293 | 190 | 172 | 150 | 319 | 151 | 118 | 213 | 114 |
| 186 | 167 | 129 | 185 | 231 | 199 | 265 | 306 | 173 | 276 |
| 291 | 231 | 205 | 220 | 246 | 239 | 186 | 299 | 233 | 208 |
| 265 | 204 | 300 | 195 | 239 | 173 | 237 | 282 | 221 | 218 |
| 197 | 215 | 213 | 290 | 146 | 232 | 305 | 184 | 149 | 267 |
| 188 | 219 | 171 | 099 | 329 | 199 | 180 | 225 | 257 | 202 |
| 189 | 207 | 792 | 327 | 201 | 300 | 206 | 199 | 299 | 153 |
| 175 | 287 | 277 | 230 | 258 | 137 | 174 | 301 | 260 | 282 |
| 211 | 212 | 284 | 214 | 283 | 139 | 223 | 212 | 207 | 224 |
| 207 | 111 | 272 | 192 | 127 | 303 | 221 | 187 | 309 | 263 |
| 203 | 176 | 233 | 239 | 176 | 218 | 193 | 243 | 236 | 275 |
| 228 | 198 | 241 | 219 | 167 | 193 | 234 | 179 | 126 | 176 |
| 279 | 178 | 275 | 260 | 191 | 174 | 235 | 338 | 242 | 238 |
| 211 | 187 | 184 | 189 | 305 | 221 | 253 | 225 | 327 | 203 |
| 195 | 158 | 156 | 185 | 170 | 271 | 160 | 188 | 165 | 218 |
| 312 | 243 | 267 | 298 | 196 | 139 | 205 | 298 | 238 | 217 |
| 145 | 201 | 313 | 230 | 185 | 166 | 147 | 223 | 271 | 133 |
| 155 | 230 | 287 | 329 | 265 | 150 | 286 | 271 | 268 | 198 |
| 214 | 231 | 163 | 335 | 198 | 270 | 187 | 174 | 163 | 201 |
| 192 | 247 | 247 | 297 | 178 | 240 | 290 | 234 | 170 | 227 |
| 230 | 353 | 170 | 159 | 236 | 181 | 230 | 240 | 212 | 242 |
| 151 | 158 | 253 | 179 | 263 | 158 | 250 | 226 | 246 | 301 |

1. Following is list of 70 villages of a Tehsil of a certain district of Punjab along with population and cultivated area of the same years select
2. a sample of 10 villages,
3. a sample of 15 villages,
4. a sample of 20 villages.

Estimate the population for three cases, estimate the standard error for each case. Compare the result and comment.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.**  **No.** | **Population** | **Cultivated**  **Area (Acres)** | **S.**  **No.** | **Population** | **Cultivated**  **Area (Acres)** |
|  | 226 | 678 |  | 904 | 760 |
|  | 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 |  | 1178 | 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 |
|  | 2155 | 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 |

1. A residential area has 5000 houses. It is required to estimate the proportion of houses with more than three persons living in them. The estimator is required to have standard error not exceeding 0.02. From other surveys, it would appear that the proportion for this is lying in the range 0.35 to 0.55. How large a sample is needed to meet the accuracy?

##### (HINT) We know that for large sample . The problem is to find p. We know that p is lying between the limits 0.35 and 0.55.P + t S.E(p).

1. Suppose the loss function in terms of money is proportional to the value of the coefficient of variation. The cost function is, C = C0 + C1n, the sample size ignoring f.p.c. will be  where Cy is coefficient of variation.

16. Following is a population of 47 villages showing area under fresh fruit in a certain distinct of Punjab. Select 10 percent sample of the villages and estimate the total area under fresh fruit along with standard error of the estimate. Compare the estimated total with actual one.

|  |  |  |  |
| --- | --- | --- | --- |
| S. No. of the **Selected Villages** | Area UnderFresh Fruit **%** | S. No. of theSelected Villages | Area UnderFresh Fruit **%** |
|  | 127.00 |  | 8.00 |
|  | 32.00 |  | 24.00 |
|  | 74.99 |  | 3.00 |
|  | 57.00 |  | 24.00 |
|  | 68.00 |  | 38.00 |
|  | 61.01 |  | 13.00 |
|  | 6.00 |  | 4.00 |
|  | 27.00 |  | 12.00 |
|  | 68.00 |  | 1.00 |
|  | 13.17 |  | 5.00 |
|  | 8.00 |  | 1.00 |
|  | 8.00 |  | 13.00 |
|  | 22.00 |  | 19.00 |
|  | 10.00 |  | 5.00 |
|  | 30.00 |  | 12.00 |
|  | 98.99 |  | 12.00 |
|  | 7.38 |  | 7.38 |
|  | 62.99 |  | 7.00 |
|  | 13.00 |  | 2.00 |
|  | 14.00 |  | 8.00 |
|  | 30.00 |  | 1.00 |
|  | 10.00 |  | 10.00 |
|  | 13.00 |  | 22.00 |
|  | 7.00 |  | **1419.91** |

##### APPENDIX- 1 ( Random Digits)

57780 97609 52482 12783 88768 12323 64967 22970 11204 37576

68327 00067 17487 49149 25894 23639 86557 04139 10756 76285

55888 82253 67464 91628 88764 43598 45481 00331 15900 97699

84910 44827 31173 44247 56573 91759 79931 26644 27048 53704

35654 53638 00563 57230 07395 10813 99194 81592 96834 21374

46381 60071 20835 43110 31842 02855 73446 24456 24268 85291

11212 06034 77313 66896 47902 63483 09924 83635 30013 61791

49703 07226 73337 49223 73312 09534 64005 79267 76590 26066

05482 30340 24606 99042 16536 14267 84084 16198 94852 44305

92947 65090 47455 90675 89921 13036 92867 04786 76776 18675

51806 61445 32437 01129 03644 70024 07629 55805 85616 59569

16383 30577 91319 67998 72423 81307 75192 80443 09651 30068

30893 85406 42369 71836 74479 68273 78133 34506 68711 58725

59790 11682 63156 10443 99033 76460 36814 36917 37232 66218

06271 74980 46094 21881 43525 16516 26393 89082 24343 57546

93325 61834 40763 81178 17507 90432 50973 35591 36930 03184

46690 08927 32962 24882 83156 58597 88267 32479 80440 41668

82041 88942 57572 34539 43812 58483 43779 42718 46798 49079

14306 04003 91186 70093 62700 99408 72236 52722 37531 24590

63471 77583 80056 59027 37031 05819 90836 19530 07138 36431

68467 17634 84211 31776 92996 75644 82043 84157 10877 12536

94308 57895 08121 07088 65080 51928 74237 00449 86625 06626

52218 32502 82195 43867 79935 34620 37386 00243 46353 44499

46586 08309 52702 85464 06670 18796 74713 81632 34056 56461

07869 80471 69139 82408 33989 44250 79597 15182 14956 70423

46719 60281 88638 26909 32415 31864 53708 60219 44482 40004

74687 71227 59716 80619 56816 73807 94150 21991 22901 74351

42731 50249 11685 54034 12710 35159 00214 19440 61539 25717

71740 29429 86822 01187 96497 25823 18415 06087 05886 11205

96746 05938 11828 47727 02522 33147 92846 15010 96725 67903

27564 81744 51909 36192 45263 33212 71808 24753 72644 74441

21895 29683 26533 14740 94286 90342 24671 52762 22051 31743

01492 40778 05988 65760 13468 31132 37106 02723 40202 15824

55846 19271 22846 80425 00235 34292 72181 24910 25245 81239

14615 75196 40313 50783 66585 39010 76796 31385 26785 66830

77848 15755 91938 81915 65312 86956 26195 61525 97406 67988

87167 03106 52876 31670 23850 13257 77510 42393 53782 32412

73018 56511 89388 73133 12074 62538 57215 23476 92150 14737

29247 67792 10593 22772 03407 24319 19525 24672 21182 10765

17412 09161 34905 44524 20124 85151 25952 81930 43536 39705

68805 19830 87973 99691 25096 41497 57562 35553 77057 06161

40551 36740 61851 76158 35441 66188 87728 66375 98049 84604

90379 06314 21897 42800 63963 44258 14381 90884 66620 14538

09466 65311 95514 51559 29960 07521 42180 86677 94240 59783

15821 25078 19388 93798 50820 88254 20504 74158 35756 42100

10328 60890 05204 30069 79630 31572 63273 13703 52954 72793

49727 08160 81650 71690 56327 06729 22495 49756 43333 34533

71118 41798 34541 76132 40522 51521 74382 06305 11956 30611

53253 23100 03743 48999 37736 92186 19108 69017 21661 17175

12206 24205 32372 46438 67981 53226 24943 68659 91924 69555

*CHAPTER 3*

# STRATIFIED AND SYSTEMATIC RANDOM SAMPLING

**3.1 INTRODUCTION**

Selection procedure of simple random sampling and its estimates have been discussed in Chapter-2. If there is much variation among the population units, simple random sampling will be less precise as the variance of sampling mean will be large. For more accurate and precise results the population is divided into different groups, called *strata*, in such a way that the population units within each stratum should be homogeneous. 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 of the individual strata. Finally estimate for each stratum are aggregated to produce an estimate for the whole population. This division reduces the variability of the strata estimator thereby achieving the higher precision. The usual purpose of stratification is to achieve an increase in the efficiency of the sample design. There are number of reasons for using this type of sampling procedure, i.e.

(i) it may increase precision by decreasing the variation within the strata, (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, selection procedure may differ from one stratum to another stratum.

The following are typical examples of stratified populations:

1. A population of school students stratified by school class by sex.
2. A population of individual taxpayers stratified by province by sex, 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.) by range of annual sales.

It will be seen that some stratification criteria are sharply defined (school class, sex, State) some admit of a 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. To begin with, the statisticians’ interest focused on the problem of allocating a sample between the strata. In the year 1926 – 1935 three papers appeared, which represented significant advances. Bowley (1926) presented a classical method of sample selection from each stratum proportion to the total number of units in that stratum. This allocation yields a *representative sample*. In Neyman (1934) it was shown 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 how to partition a population into a strata was dealt especially witha principle of equi-partitionand 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 this Chapter other than those, mentioned in the Chapter-2.

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

N = N1 + N2 + N3 + ---+ Nh + --- + Nk, where N is the number of population units in the hth stratum, similarly

n = n1 + n2 + n3 + ---+ 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.

**3.3 EXPECTATION OF SAMPLE MEAN**

**THEOREM (3.1)**

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

## PROOF

Taking expectation of the same mean

◊ as .

Similarly it can be proved that 

**3.4. Variance and Unbiased Variance Estimator**

In this Section the variance formula of  for

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

has been derived.

**3.4.1. Arbitrary Allocation**

In this type of allocation the population is divided into different strata such that the units within each stratum must be homogeneous and a sample is allocated *arbitrarily*among 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  of stratified random sampling for finite population sampling is

 **(3.4.1)**

**PROOF**

We know that sample mean for stratified random sampling is



Since the variance of sum of independent random variables is the sum of the variance of these variables, it follows

 **(3.4.2)**

[The covariance term will vanish as samples are drawn independently] Using the concept of simple random sampling, the variance of sample mean of hth stratum can be written as:

 (3.4.3)

Using (3.4.3) in (3.4.2) we get:

 **(3.4.1)**

(3.4.1) may be written as:

 **(3.4.4)**

If , (3.4.4) will be

 **(3.4.5)**

If N is large, 1/N ~ 0 then (3.4.5) is

 **(3.4.6)**

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

 **(3.4.7)**

For large N

 **(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 or, if some strata are more expensive to sample from the others, a fixed survey cost consider first 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 i.e.

 **(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 or similar classification. 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 in stratified random sampling if  is

 **(3.4.10)**

**PROOF**

Substituting the value of  from (3.4.9) in (3.4.1) and on simplification



ignoring fpc (3.4.8) will be

 **(3.4.11)**

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

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

**(i) Minimum variance for fixed cost**

**Theorem (3.4)**

In stratified random sampling  will be minimum subject to the cost when nh 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, C0 = overhead cost, and Ch = cost per unit in the hth stratum.

The objective is to choose nh, by minimizing , subject to given cost. A concept of Lagrange’s multiplier may be introduced such that

 **(3.4.13)**

From (3.4.3) and (3.4.13) we have



Finding the partial differentiation w.r.t. nh and equating to zero we have



or

 **(3.4.14)**

Summing (3.4.14) over the strata

 **(3.4.15)**

Eliminating  from (3.4.14) and (3.4.15) we get

 **(3.4.16)**

If the cost is constant for all the strata i.e. C1 = C2 = … = C then the cost function will be

C = C0 + nC **(3.4.17)**

The optimum sample size for the hth stratum from (3.4.12) will be

 **(3.4.18)**

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 (3.4.16) in (3.4.1).

 **(3.4.19)**

Ignoring fpc (3.4.19) becomes

 **(3.4.20)**

If the cost is constant then (3.4.19) will be

 **(3.4.21)**

For large N, (3.4.21) becomes,

 **(3.4.22)**

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

 **(3.4.23)**

This is the minimum sample size for estimating the mean with fixed variance. A special case of (3.4.23) may be obtained when the cost is the constant for each stratum i.e. (3.4.23) reduces to

 **(3.4.24)**

The main difficulty of 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 from highly skewed population. An example would be a 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.

**(iii) Estimating the sample size from proportional allocation**



so

 **(3.4.25)**

**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), (3.4.19) and (3.4.21) 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 for different allocations themselves and with simple random sampling.

* + 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)**

(3.6.3) may be written as

 **(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 as different from each other as possible.

This difference is non negative, being  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 hence the aim is to minimize the variance of  for a given sample size. The method of undetermined multipliers can be used to use that in this case the nh must be proportional to NhSh and the further reduction in variance as compared with proportional sample may be obtain as If term 1/Nh in all the strata is negligible then, from (3.4.10) and (3.4.25)



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.





 **(3.6.6)**

This is positive, hence



**3.6.3 Optimum Allocation and Simple Random Sampling**

From (3.6.3) and (3.6.8) 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

 **(3.6.8)**

**3.6.4. Simple Random Sampling and Arbitrary Allocation**

 **(3.6.9)**

which is also positive, hence  if nh is taken proportional to Wh as the first term of (3.6.11) reduces to zero. Also if nh is proportional to then



if 1/Nh is ignored

 **(3.6.10)**

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

Given n, the sample size for optimum allocation is

 **(3.6.11)**

with variance

 **(3.4.21)**

Suppose nh is the sample size used in hth stratum for arbitrary allocation then

 **(3.4.3)**

Increase in variance caused by imperfect allocation

 **(3.6.12)**

From (3.6.11)

 **(3.6.13)**

Using (3.6.13) in the first term of (3.6.12), we get



Since  we then have



 **(3.6.14)**

Now

 = 

Hence

****  **(3.6.15)**

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

 **(3.6.16)**

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

 **(3.6.17)**

as  is little less.

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 let we have 7 strata with  are known

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| 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 variance and standard deviation of each stratum is given.

|  |  |  |  |
| --- | --- | --- | --- |
| 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 has been selected from 468. On the basis of above information estimate the sample size for proportional and optimum allocation.  and .

**SOLUTION:**

For proportional allocation  whereas for optimum allocation . The calculation are given as:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |





If fpc is ignored

, 

, 

If fpc is ignored

, 



, .

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 | standard division |
| 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 variances of the estimates of the total number of cigarettes smoked per day and average number of cigarettes smoked per day, using

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

Check the difference between these variances using the standard formulae.

(Source New South Wales University).

**SOLUTION**

**(i) Simple Random Sampling**

 =  =  = 0.707698

**(ii) Proportional Allocation**

nh = 

Therefore n1­ = 193, n2 = 255 and n3 = 353

 = 

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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**

nh = 

Therefore n1 = 193, n2 = 255 and n3 = 352

 = 0.521081

**(iv)** may also be calculated by (3.6.3)

= =0.62028 + 0.0877 = 0.70835.

**Example 3.3**

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

**SOLUTION**

The Lagrangian may be put as



Finding the partial differentiation w.r.t. nh and equating to zero

 **(A)**

Summing over nh

 **(B)**

Comparing (A) and (B)



Hence nh is proportional to 

**3.7 STRATIFIED SAMPLING FOR PROPORTION.**

Like simple random sampling, proportion may also be used in stratified random sampling.

An unbiased estimator for population proportion appropriate to 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) of Section (3.4) to find the variance and an unbiased variance estimator.

The variance of pst for arbitrary allocation is



We know that

 **(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 in to K strata each stratum have N1, N2, …. 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 (3.4.8) we obtain



For large N, (1-n/N) is taken as unity; hence

 **(3.7.5)**

Similarly putting the value of S2h in (3.4.20) we get

 **(3.7.6)**

For large N, second term of the above expression is approximately zero.

 **(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 [(3.6.3) and 3.6.5)].

 **(3.7.9)**

and

 **(3.7.10)**

**Example 3.6**

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 as

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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





The reduction in variance due to optimum allocation is



**Example 3.7**

2300 households in a city area was divided into 5 stratum on the basis of their monthly income. Simple random sampling technique was applied to select a sample within each stratum and information were 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 household rent in the population is



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

The variance of pst



 = 9522

 = 97.5807

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

Sometimes in a survey it is useful or of a interest to find out how stratification is effective as compared to simple random sampling. In comparing the precision of stratified with simple random (un-stratified) 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 un-stratified 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. Without going into details of the algebraic manipulation the variance estimator of  given by J.N.K. Rao (1962) is reproduced as:

 **(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)**

A finite population of size N with parameters  has two strata of size N1 and N2 where the first stratum has all zero values, so  then



**SOLUTION**



We know that



since 





since N = N1 + N2 therefore



**Example 3.10**

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 tries 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

 using (3.8.1) and ignoring f.p.c. is

 = 

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%

**3.12 QUOTA SAMPLING**

Various agencies in a country conduct certain surveys to elicite public opinions, political views, reaction to a new governmental policy, consumers’ attitude to a marketable products etc. many commercial organizations also conduct nation-wide inquiries for assessment of public attitude towards a particular aspects of national activity. Such investigations are popularly made on the bias of a selection; of a so-called representative sample that gives nice picture about the prevalent opinion of a certain class of people. In stratified random sampling the strata are chosen so far as the units in the to population are homogeneous whereas in quota sampling strata are chosen to ensure as representativeness of population units in the sample in order to obtain as good estimates in the specific stratum rather than the efficiency of the estimates. In quota sampling strata are determined with respect to a particular characteristic for obtaining a representative estimate whereas in stratified random sampling strata are constructed to increase the efficiency of the estimate. We may need to obtain public opinion about political issue from a certain age group or some professional group.

In quota sampling elements of interview’s choice in the selection of sample unit is predetermined, but it does not exist in stratified sampling. In fact quota sampling should not be compared to stratified sampling, though there exist some design affinity. Interviewers fill the questionnaires by person of a particular category whom they meet in street without random process. The results of the quota sampling cannot be generalized.

This process has many disadvantages (i) its scope is limited (ii) interviewers’ choice are included in the selection procedure making estimates highly biased (iii) non respondent cannot be estimated (iv) the result of the survey cannot be used in other survey.

In view of these disadvantages the quota sampling has many uses. It can produce good results without much cost in case where probability sampling may not give proper results in a certain period of time. If probability sampling has to be applied in opinion polls from a certain age group, it will be very costly and time consuming process because to interview persons of age 25-30 years of married females of a income group 100-200 Dinars, the female will be highly sophisticated. In such cases any probability sampling will fail to yield good results, whereas quota sampling will give quick, and sufficient good estimates. Quota sampling has been considered in detail by Stephen and McCarthy (1958).

**SYSTEMATIC SAMPLING**

**3.13 Introduction**

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 selection only the first unit is to be selected at random and the remaining units are automatically determined by the skip interval. 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 would be divided into n groups each containing k units. If rth (say) unit is selected at random from the first group of *k* units, then (*k+r*)th unit, (2*k+r*)th unit would be 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***.* This procedure of selecting the sample is called **SYSTEMATIC SAMPLING** or **Systematic Random Sampling.**

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 4.1

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Sample Composition | Probability  P(S) | Sample  Mean |
| 1 | 1, k+1, 2k+1, …, (i-1)k+1, … (n-1)k+1 | 1/k |  |
| 2 | 2, k+2, 2k+2, …, (i-1)k+2, … (n-1)k+2 | 1/k |  |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| r | r, k+r, 2k+r, (i-1)k+r, … (n-1)k+r | 1/k |  |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| ‘ | ‘ ‘ ‘ ‘ ‘ | ‘ | ‘ |
| k | k, 2k, 3k ,… ,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 a 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 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 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 psuedo-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 s2 calculated from (as given in Chapter 2)

  (2.5.6)

will be much 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.

 (4.1.1)

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 (4.1.1). 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.

**4.2. Expectation of Sample Mean.**

**Case (i): *N* = *nk***

**Theorem (4.1):** 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. 

We know (see table 4.1) that

, (4.2.1)

and

, (4.2.2)

then

. (4.2.3)

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

 (4.2.4)

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

We know that

 (4.2.2)

Substituting  in (4.2.4), we get



Hence  is an unbiased estimator of. ◊

#### Case (ii): N ≠ nk

If N/n is not an integer, the sample does not remain fixed as it depends on the random start. This is the limitation of systematic sampling. In these situations unbiased estimates are obtained by modified methods.

1. Suppose we have a population of 7 units, say Y1, Y­2, …, Y7, with k = 3. If the random start is the first number the sample units will be Y1, Y4, and Y7, but for random start 2nd and 3rd the sample will be Y2, Y5 and Y3, Y6 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 kth unit both backward and forward. If in the above population the random start from 1 to 7 is 4 then the sample is Y1, Y4, and Y7 with probability 3/7, if the random start is second or third unit the samples will be Y2, Y5 and Y3, Y6 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 will be

 (4.2.5)

**4.3. Variance and Variance Estimator of Sample Mean**

**Theorem (4.2):** In systematic sampling of size n, selected from a population of N units [ integer] the variance of is

 (4.3.1)

or

 (4.3.2)

where

 (4.3.3)

**Proof:** Since there are k possible samples,  is the mean of rth systematic sample, therefore,





To prove (4.2.2) let S2 is,



Adding and subtracting  in the right hand side of above expression and on simplification

 (4.3.4)

Using (4.3.3), in (4.3.4) we have

 (4.3.5)

Hence from (4.3.4) and (4.3.1), we get

◊

The total variance S2 may be splitted in terms of between variance  and within variance  from (4.3.4)



From analysis of variance technique we can write

 (4.3.6)

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

or

 (4.3.7)

**Theorem (4.3):** 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

 (4.3.2)

Comparing (2.4.1) and (4.3.2) we get



if  (4.3.8)

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



Substituting the value of  in the above expression we get





 (4.3.9)

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



 (4.3.10)

[Here the numerator is averaged over all kn(n-1)/2 distinct pairs and the denominator is (N-1)S2/N]

Since S2 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 should be heterogeneous. So the variance of systematic samples does not depend on S2 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 is [using (4.3.9) nd (2.4.1)]

 (4.3.11)

It is clear that,





and



If ρ = 1 then = (N-1)/ (k-1).

Further suppose  is the mean of the ith stratum i.e.

 (4.3.12)

The stratum mean square within ith stratum is

 (4.3.13)

The pooled mean square between units within stratum is

 (4.3.14)

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

 (4.3.15)

Using (4.3.14)

 (4.3.16)

The variance of  may be written as:









, (4.3.17)

where



Comparing (4.3.16) and (4.3.17), we have

 (4.3.18)

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







**4.3.1. Variance Estimator.**

Since the systematic sampling procedure does not ensure the inclusion of each  pairs of units of the universe 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 estimate 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. A variance estimate of sample mean and estimated total  may be suggested as

 (4.3.19)

and

 (4.3.20)

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 S2. 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 (thi being th part of the population).

**Example (4.1):** 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 n1 = n2 = n3 = n4 = 5 and n5 = 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 (4.2):** 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.

**Table 4.2**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |

**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. The population and cultivated area are given against each village as:

**Table 3**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sr.No | No.Villag-es | Population | Cultivated  Area | Population  (yi – yi+1)2 | Cultivated Area  (yi – yi+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

|  |  |  |
| --- | --- | --- |
|  | 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:** 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.

**Table 4.4**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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** |

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 |

**4.4. Relative Efficiency of Systematic Versus Simple Random Sampling**

The relative efficiency of systematic sampling compared to simple random sampling with and without replacement respectively is as





**4.5. 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 kth unit is selected in a cyclical manner till required sample of size n is obtained. This procedure is 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 i.e.

r + ik if r + ik ≤ N

and

r + ik – N if r + ik > N

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

It can be easily seen that probability of selection is equal and is 1/N. An unbiased ness under circular systematic is shown as by an artificial example. Let N = 14, (1,2,3,…14)   
n = 5 and k = 3. All possible samples with different random starts [1 to N] 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 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 n1 = n2 = 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 (Σyhi) | Σy2hi |
| 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 as 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 attitude 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 its 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 nh under optimum and proportional allocation when n = 100. Also find the  when cost is given.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stratum | Nh | S2h |  | 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 n1/n and n2/n which minimize the total cost   
   C = c1n1 + C2n2 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 | Sh  per-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. The variate y has rectangular distribution in the interval (a, a + d). The interval is divided into K sub-intervals which form K-strata of equal size. From each stratum, a simple random sample of n/K units is drawn. Let V1 and V2 be the variance of stratified and un-stratified samples of size n respectively. Prove that V1/V2 = 1/K2.
3. The right triangular distribution f(y) = 2(1-y), 0 < y < 1, is divided into two strata at the point a show that   
     
    W1 = a(2 – a) W2 = (1 – a)2   
     
    
4. With two strata, a sampler would like to have n1 = n2, instead of using the values given by Neyman allocation. If ,  denote the variance given by the n1 = n2 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 = W1S1/S2S2 **]**

1. A sampler has two strata with relative sizes W1 and W2. He believes that S1 and S2 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 = C1n1 + C2n2, 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. A 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. Suppose 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. Suppose 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

1. the random start is 14.
2. random start is 12.  
   Compare the variances of the two samples.
   1. From the data given in Example 4.2,
3. draw five circular systematic samples of size 7 each, from a rearranged frame.
4. from each of the five samples, estimate the total cultivated area in the tehsil.
5. obtain a single combined estimate from the five sample estimates. Also, calculate the standard error of this combined estimate.
   1. A 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.
   1. 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.
   1. 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.

*CHAPTER 4*

**SINGLE STAGE CLUSTER SAMPLING**

**4.1. INTRODUCTION.**

Historically speaking the term **cluster** was first used 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 will be selected from the villages, each selected villages 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 than other sampling designs.

Since 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 be less. The efficiency of clustering sampling depends on the size of the cluster. If the size is large the efficiency will be decreased, if the size is small and number of clusters are more it will be increased. In a household survey, the households if selected randomly and independently will represent the whole population and thereby provide better estimate than cluster sampling. Experiences have shown that cluster sampling procedure will be 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 may 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 must be homogeneous. 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 sample is selected within each group and selection is independent within the groups Simple random sampling, systematic random sampling or probability proportional to size sampling (if the measure of size is know) may be used while selecting a sample.

Cluster should be formed in such a way that the variances within the cluster should as apart as possible whereas in stratified sampling variances within the strata should be as close as possible.

**4.2 CLUSTER OF EQUAL SIZES.**

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C l u s t e r s | | | | | | | | |
| ↓j i→ | 1 | 2 | 3 | ......... | i | ......... | N |  |
| 1 | Y11 | Y21 | Y31 | ......... | Yi1 | ......... | YN1 |  |
| 2 | Y12 | Y21 | Y32 | ......... | Yi2 | ......... | YN2 |  |
| 3 | Y13 | Y21 | Y33 | ......... | Yi3 | ......... | YN3 |  |
| . | . | . | . | ......... | . | ......... | . |  |
| . | . | . | . | ......... | . | ......... | . |  |
| . | . | . | . | ......... | . | ......... | . |  |
| j | Y1j | Y2j | Y3j | ......... | Yij | ......... | YNj |  |
| . | . | . | . | ......... | . | ......... | . |  |
| . | . | . | . | ......... | . | ......... | . |  |
| . | . | . | . | ......... | . | ......... | . |  |
| M | Y1M | Y2M | Y3M | ......... | YiM | ......... | YNM |  |
| Mean | ‾Y1 | ‾Y2 | ‾Y3 | ......... | ‾Yi | ......... | ‾YN |  |

Where Yij denotes the jth element of the ith cluster (i = 1, 2, 3, ......... N)

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

Then we can define the notation

Population mean

Mean of the ith cluster 

Between cluster variance 

Within cluster variance 

Total variance  (4.2.1)

Total variance may be split into between and within variance as





 (4.2.2)

or  (4.2.3)  (4.2.4)

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 ith cluster 



 (5.2.5)

Total variance may be written analogous to (5.2.2) as

 (5.2.6)

If n and M are sufficiently large (5.2.5) becomes as

 (5.2.7)

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 |  |  |

**5.3. EXPECTATION**

**THEOREM (5.1)**

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

 (5.3.1)

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

Taking expectation of (5.3.1) we get



 (5.3.2)

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.

 (5.3.3)

**5.4 VARIANCE AND UNBIASED VARIANCE ESTIMATOR**

**theorem (5.2)**

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

 (5.4.1)

# 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 (5.4.1)

This may be put as (using )

 (5.4.2)

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 (5.2.2) with re-arrangement as

 (5.4.3)

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

 (5.4.4)

**THEOREM (5.3)**

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

, (5.4.5)

# PROOF

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

 (5.4.6)

 (5.4.6)

The variance of estimated total  is

 (5.4.7)

**THEOREM (5.4)**

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 is drawn from a population of NM elements, the cluster sample will be more precise than simple random sample provided within variance of the clusters is more than over all variance.

# PROOF

From (5.4.1) and (5.4.6) we get  (5.4.8)

 (5.4.9)

From (5.2.2) we get

 (5.4.10)

From (5.4.10) in (5.4.9) we get

  (5.4.11)

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.

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

In numerous cases, a general measure is required which gives some indication as to the accuracy of clustering sampling and to the effect of cluster size has on the sampling scheme. Many people 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.

 (5.4.12)

[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)S2/NM].

The variance of sample mean  may also be expressed in terms of intra-class correlation coefficient. For that let



 (5.4.13)

Using (5.2.1) and (5.4.12) in (5.4.13), we have



or

 (5.4.14)

**Theorem (5.5)**

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

 (5.4.15)

# PROOF

We know that

 (5.4.1)

Using (5.4.14) in (5.4.1), we get

 (5.4.15)

If M and N are large than (5.4.15) is

 (5.4.16)

Ignoring f.p.c. (5.4.16) becomes

 (5.4.17)

Variance formula for the estimation of total may be written as

 (5.4.18)

It appears that  not only depends on samples size n, but size of the cluster M, and intra-class correlation coefficient) are identical and we say both procedure are equally good. If M > 1 and ρ is positive. In this case cluster sampling procedure is less precise. 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 relative efficiency of cluster sampling compared with simple random sampling i.e. comparing (2.4.1) and (5.4.16), we get

 (5.4.19)

Also relative efficiency of cluster sampling and random sampling when nM elements directly selected from NM elements, comparing (5.4.5) and (5.4.17), we get

 (5.4.20)

If M = 1 then (5.4.19) and (5.4.20) are identical and

.

**5.4.2 Calculation of ρ.**

Experience shows that the magnitude of the coefficient of intraclass correlation usually decreases as the size of the cluster increases with the rate of the latter being faster. The knowledge of the 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 way for the calculation of ρ follows as:

(i) We know that

 (5.4.14)

We know that



Using this in (5.4.14)

 5.2.21) or

 (5.4.22)

or  (5.4.23)

(ii) or alternatively, we know that

 (5.2.2)

Substituting the value of  from (5.4.21) in (5.2.2) and on simplification

 (5.2.24) or  (5.4.25)

**5.2.2 Estimated Variance**

An unbiased estimator of (5.4.1) may be written in a simple way as

 (5.4.26)

The variance of estimated total  is

 (5.4.27)

where

. (5.4.28)

An unbiased variance estimator of  is

 (5.4.29)

 (5.4.30)

Infant 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 and the effect of the clustering is to 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, ρ = 1 and the effect of clustering is to 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, ρ = – 1/(M–1) and the variation is zero. This makes sense, in that a sample of only one cluster then supplies perfect information about the entire population.
3. When the clusters are random aggregations of population units, ρ = 0 and the variance of a clustered sample of n units is identical with the variance of a clustered sample of n units is identical with the variance of a simple random sample of n units.

# EXAMPLE 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:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| c | **1st Tree** | **2nd Tree** | **3rd Tree** | **4th Tree** | **Total** |
| 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 with simple random sampling.

# SOLUTION

There are 420 trees and each cluster is having 4 trees.

Therefore population size , a sample of 15 cluster was selected.

To calculate we proceeds as follows:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cr |  |  |  |  |  |  |  |
| 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 |

Following calculations are made



B.S.S (Between clusters) = 

T.S.S 



W.S.S (within clusters) 



These calculations may be put 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



Relative efficiency of cluster sampling with that of simple random sampling.

B.S.S 

T.S.S 

W.S.S 



# EXAMPLE

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:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 be found



Substituting the value from the table it comes out to be



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





ρ may also be estimated using (5.4.28) as



**5.5 DETERMINATION OF CLUSTER SIZE**

It has been already pointed out that variance of sample mean in cluster sampling depends on the number of cluster, S2, and intra-class correlation coefficient ρ. We have also seen that p depends on the size of the cluster that varies from population to population. In this section the optimum level of cluster will be 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. Hanif and Ahmad (1977) have also determined the optimum level of cluster size for multipurpose sample surveys in Libya.

The problem is to find the optimum value of cluster when the cost is involved. Cost of the survey depends both on the size of the cluster and number of clusters.

Let us take a simple cost function.

 (5.5.1)

C = total cost, C0 = over head cost

C1 = cost per cluster C2 = cost per element in each cluster

Let the variance of sample mean  is

 (5.4.18)

The objective is to find M by minimizing  subject to cost given in (5.5.1). The value of n from (5.5.1) is

 (5.5.2)

Substituting the value of n from (5.5.2) in (5.4.18) the function is



Finding the partial differentiation with respect to M and equating to zero



On simplification we obtain

 (5.5.3)

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

For ρ = 0; M → ∞, for ρ = 1; M → 0

For ρ = 0.001;  and for ρ = 0.1; 

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

If the number of clusters are increased then the cost function given by Jessen (1942) is

 (5.5.4)

where

C = total cost,

C1 = cost of interview and cost of travel from one element

to another element, and

C2 = cost of travel between clusters,

Let the variance of  be

 (5.4.18)

differentiating with respect to M

 (5.5.5)

also differentiating (5.5.4) with respect to M and on simplification

 (5.5.6)

Substituting dn/dM in (5.5.5) and on simplification, when dV/dM = 0

 (5.5.7)

Solving (5.5.4) as a quadratic in 

 (5.5.8)

From (5.5.7) and (5.5.8) we have



or



or

 (5.5.9)

Jessan (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].

# EXAMPLE 3

For a survey on birth rate, given that the total cost, neglecting over heads, 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

C1 = 2 man-day = Rs. 20 C2 = 1/40 man-day = 0.25

C1/C2 = 80, ρ = 0.001

The optimum size of cluster from (5.5.3) is

M = 

Total cost (neglecting overhead)

C = C1n + nMC2

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 C1/C2 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

* 
* n2 – 2900n + (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 as

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

**5.6 CLUSTER SAMPLING FOR PROPORTION.**

If Pi is the population proportion of the ith cluster, then it may be defined as

Pi = Ai/M (5.6.1)

And if a sample of n clusters is taken the pi denotes the proportion of the ith cluster

pi = ai/M (5.6.2)

The unbiased estimator of population proportion may be defined as

 (5.6.3)

The variance of pc may be written directly (using the concept of Chap 2) from (5.4.1) when 

 (5.6.4)

and an unbiased estimator of (5.6.4) is

 (5.6.5)

**5.7 CLUSTER UNEQUAL SIZES**

So far we have discussed only clusters of equal sizes. In actual practical situations clusters i.e. block, households etc. ordinarily used in the design of sample survey do not contain the same number of elements. Let the population is divided into N cluster each having M1, M2, ......, MN elements, then the population mean is

 (5.7.1)

,

where



When the clusters are unequal in 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 simple random sampling procedure. 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:

(a)  (5.7.2)

where

 (5.7.3)

which is clearly unbiased.

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

 (5.7.4)

and an unbiased estimator of (5.7.4) is

 (5.7.5)

(5.7.5) may be written as

 (5.7.6)

(b)  (5.7.7)

 (5.7.8)

The variance and unbiased variance are

 (5.7.9)

and

 (5.7.10)

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

(c) Ratio-to-size estimate

Let

 (5.7.11)

Ratio-to-size estimate could be

; (5.7.12)

where Mi are known, Mi are taken as bench mark variable.

We know that ,  may be written using the concept of ratio (given in Chapter 6).

 (5.7.13)

and approximation unbiased estimator for (5.7.13) is

 (5.7.14)

# EXAMPLE

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 |

# 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 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 types of sampling units as:

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.

We can write all these values in a 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 sample 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 is 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 C1 increases or C2 decreases.
3. Suppose in a study on cluster sampling, of n cluster 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.

*CHAPTER-5*

# MULTISTAGE CLUSTER SAMPLING

**12.1 INTRODUCTION**

Simple random sampling and stratified random sampling schemes described before may be considered as single stage sampling schemes. In a single stage scheme, a sample is drawn from a population and information are obtained from the selected sampling units. In multistage sampling, a population is divided into certain large units and a sample of large units is further subdivided into smaller units, and a sample of these units is selected from each of the selected large units. Kendall and Buckland (\_\_\_\_) in Dictionary of Statistical Terms define a multistage sample as one “which is selected by stages, the sample units at each stage being sub-sampled from the (larger) units chosen at the previous stage”. In other words multistage sampling is the term applied to that kind of sampling in which selection is carried by stages where the sampling units at each stage are sub-sampled from the larger units chosen at the previous stage. Thus, a municipality may be divided into certain number of zones and a number of zones are selected randomly as **first-stage** sampling units. Within each zone drawn in the sample at random, a number of schools are chosen at random as **second-stage** sampling units. Within each school drawn in the sample, a sample of students can be randomly selected as **third-stage** sampling units. This is an example of a three-stage sampling, where zone is **first-stage**, school is **second -stage** and students selected is the **third-stage(tertiary sampling units)** (last stage) Here students drawn within schools composed the sample to be analyzed. It is often used in combination with area sampling and cluster sampling. In-fact at the first stage the entire population is divided into primary stage units (P.S.U’s) and a selection is made from them. At each successive stage, smaller sampling units are defined within those selected at the previous stage, and a further selection is made within each of them.

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

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

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

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

**12.2 NOTATION**

|  |  |  |
| --- | --- | --- |
| N | = | The number of first stage units in the population. |
| Mi | = | The number of second stage units in the ith first stage unit. |
| Qij | = | The number of third stage units in the ith first stage and jth second stage. |
| n | = | The number of first stage units in the sample. |
| mi | = | The number of second stage units to be selected from the ith first stage sample unit. |
|  |  |  |
| qij | = | The number of third stage units to be selected ith first and jth second stage sample unit. |
|  |  |  |
| Yi | = | Population values for the ith first stage unit. |
| Yij | = | Population values for the jth second stage units the ith first stage. |
| Yijk | = | Population for the kth third stage with the ijth second stage unit. |

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

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

**12.3 VARIANCE ESTIMATION FOR MULTISTAGE SAMPLING**

A basic principle of multistage sampling is that when selection and estimation take place independently at various stages, the variances of an unbiased estimator which arise from each of these different stages can be added. In particular the total variance of such a multistage estimator is equal to the variance arising from the first stage plus that arising from subsequent stages. Formerly this may be written as

 (12.3.1)

where E1 denotes the expectation and V1 is the variance over all first stage samples, E2 denotes the expectation over all second and subsequent stage samples, and  is the conditional variance of y' subject to the selection of a particular first stage sample. With an obvious extension of this notation three stages may be written as

  (12.3.2)

and for k stages sampling as

 (12.3.3)

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

* 1. **TWO STAGE SAMPLING**

In this section two stage sampling for various selection procedures will be discussed. In each case estimation of population total, its variance and variance estimators have been derived.

**12.4.1 Equal Probability at each Stage**

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

  (12.4.1)

 as 

, as ,

where yij is the value of jth second stage unit from the ith selected primary stage unit. This technique is know as two stage sampling or sub-sampling.

#### THEOREM 12.1

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

**PROOF:**

The unbiasedness may be proved as; (The expectation will be involved at two stages).

 





Taking first stage expectation 

.◇

#### THEOREM 12.2

The variance of (12.4.1) is

 (12.4.2)

where

.

#### PROOF:

We know that



 (12.4.3)

After substituting the value of  from (12.4.1) in (12.4.3) and squaring we get



where 

Let us first take the third term and take second stage expectation, which will be

 as 

We left with (12.4.5)

Now we left with the first two terms

The second term will be

, (Primary stage variance) (12.4.6)

where 

and the first term of (12.4.5)





The second term  it may be proved as



we left with

= (12.4.8)









 (12.4.9)



Putting (12.4.6) and (12.4.9) in (12.4.5) we get (12.4.2) ◇

(12.4.2) may be written as

, (12.4.10)

where  and 

If  all the clusters are equal in size and sampling fraction in each cluster at the second stage in uniform then (12.4.10) is

, (12.4.11)

where .

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

The variance for sample mean  may be written from (12.4.10) as

. (12.4.12)

If all the clusters are equal in sizes then variance of sample mean is

. (12.4.13)

 may alternatively be derived by using (12.3.1). For this we have

 (12.4.14)

and

.12.4.15)

Substituting (12.4.14) and (12.4.15) in (12.3.1) we obtain (12.4.2). ◇

**12.4.1 Unbiased Variance Estimator**

#### THEOREM 12.3

An unbiased variance estimator of Var(y') is

, (12.4.16)

where

,

and

.

#### PROOF

Starting with the expectation of  we have

 

 (12.4.17)

Now the first term

 

.

Taking the first stage expectation

 (12.4.18)

and

 

 (12.4.19)

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

Taking the first stage expectation of (12.4.19)





 (12.4.20)

Substituting (12.4.18) and (12.4.20) in (12.4.17)

 2.4.21)

Since in the first term of (12.4.2) is a multiple factor of  so we also multiply (12.4.21) by the same factor and on simplification



or



Since



Therefore

 is an unbiased estimator of (12.4.2). ◇

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

 *CHAPTER-5*

# MULTISTAGE CLUSTER SAMPLING

**12.1 INTRODUCTION**

Simple random sampling and stratifie√d random sampling schemes described before may be considered as single stage sampling schemes. In a single stage scheme, a sample is drawn from a population and information are obtained from the selected sampling units. In multistage sampling, a population is divided into certain large units and a sample of large units is further subdivided into smaller units, and a sample of these units is selected from each of the selected large units. Kendall and Buckland (\_\_\_\_) in Dictionary of Statistical Terms define a multistage sample as one “which is selected by stages, the sample units at each stage being sub-sampled from the (larger) units chosen at the previous stage”. In other words multistage sampling is the term applied to that kind of sampling in which selection is carried by stages where the sampling units at each stage are sub-sampled from the larger units chosen at the previous stage. Thus, a municipality may be divided into certain number of zones and a number of zones are selected randomly as **first-stage** sampling units. Within each zone drawn in the sample at random, a number of schools are chosen at random as **second-stage** sampling units. Within each school drawn in the sample, a sample of students can be randomly selected as **third-stage** sampling units. This is an example of a three-stage sampling, where zone is **first-stage**, school is **second -stage** and students selected is the **third-stage(tertiary sampling units)** (last stage) Here students drawn within schools composed the sample to be analyzed. It is often used in combination with area sampling and cluster sampling. In-fact at the first stage the entire population is divided into primary stage units (P.S.U’s) and a selection is made from them. At each successive stage, smaller sampling units are defined within those selected at the previous stage, and a further selection is made within each of them.

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

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

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

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

**12.2 NOTATION**

|  |  |  |
| --- | --- | --- |
| N | = | The number of first stage units in the population. |
| Mi | = | The number of second stage units in the ith first stage unit. |
| Qij | = | The number of third stage units in the ith first stage and jth second stage. |
| n | = | The number of first stage units in the sample. |
| mi | = | The number of second stage units to be selected from the ith first stage sample unit. |
|  |  |  |
| qij | = | The number of third stage units to be selected ith first and jth second stage sample unit. |
|  |  |  |
| Yi | = | Population values for the ith first stage unit. |
| Yij | = | Population values for the jth second stage units the ith first stage. |
| Yijk | = | Population for the kth third stage with the ijth second stage unit. |

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

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

**12.3 VARIANCE ESTIMATION FOR MULTISTAGE SAMPLING**

A basic principle of multistage sampling is that when selection and estimation take place independently at various stages, the variances of an unbiased estimator which arise from each of these different stages can be added. In particular the total variance of such a multistage estimator is equal to the variance arising from the first stage plus that arising from subsequent stages. Formerly this may be written as

 (12.3.1)

where E1 denotes the expectation and V1 is the variance over all first stage samples, E2 denotes the expectation over all second and subsequent stage samples, and  is the conditional variance of y' subject to the selection of a particular first stage sample. With an obvious extension of this notation three stages may be written as

  (12.3.2)

and for k stages sampling as

 (12.3.3)

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

* 1. **TWO STAGE SAMPLING**

In this section two stage sampling for various selection procedures will be discussed. In each case estimation of population total, its variance and variance estimators have been derived.

**12.4.1 Equal Probability at each Stage**

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

  (12.4.1)

 as 

, as ,

where yij is the value of jth second stage unit from the ith selected primary stage unit. This technique is know as two stage sampling or sub-sampling.

#### THEOREM 12.1

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

**PROOF:**

The unbiasedness may be proved as; (The expectation will be involved at two stages).

 





Taking first stage expectation 

.◇

#### THEOREM 12.2

The variance of (12.4.1) is

 (12.4.2)

where

.

#### PROOF:

We know that



 (12.4.3)

After substituting the value of  from (12.4.1) in (12.4.3) and squaring we get



where 

Let us first take the third term and take second stage expectation, which will be

 as 

We left with (12.4.5)

Now we left with the first two terms

The second term will be

, (Primary stage variance) (12.4.6)

where 

and the first term of (12.4.5)





The second term  it may be proved as



we left with

= (12.4.8)









 (12.4.9)



Putting (12.4.6) and (12.4.9) in (12.4.5) we get (12.4.2) ◇

(12.4.2) may be written as

, (12.4.10)

where  and 

If  all the clusters are equal in size and sampling fraction in each cluster at the second stage in uniform then (12.4.10) is

, (12.4.11)

where .

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

The variance for sample mean  may be written from (12.4.10) as

. (12.4.12)

If all the clusters are equal in sizes then variance of sample mean is

. (12.4.13)

 may alternatively be derived by using (12.3.1). For this we have

 (12.4.14)

and

.12.4.15)

Substituting (12.4.14) and (12.4.15) in (12.3.1) we obtain (12.4.2). ◇

**12.4.1 Unbiased Variance Estimator**

#### THEOREM 12.3

An unbiased variance estimator of Var(y') is

, (12.4.16)

where

,

and

.

#### PROOF

Starting with the expectation of  we have

 

 (12.4.17)

Now the first term

 

.

Taking the first stage expectation

 (12.4.18)

and

 

 (12.4.19)

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

Taking the first stage expectation of (12.4.19)





 (12.4.20)

Substituting (12.4.18) and (12.4.20) in (12.4.17)

 2.4.21)

Since in the first term of (12.4.2) is a multiple factor of  so we also multiply (12.4.21) by the same factor and on simplification



or



Since



Therefore

 is an unbiased estimator of (12.4.2). ◇

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

 (12.4.22)