Directorate of Distance Education

UNIVERSITY OF JAMMU
JAMMU

SELF LEARNING MATERIAL
B.A SEMESTER -IV

SUBJECT : STATISTICS **UNIT : I - IV**
COURSE CODE : ST 404 **LESSON NO. 1-14**

DR. HINA S. ABROL
Course Co-ordinator

B.A. SEMESTER -IV

STATISTICS

Content Editing & Proof Reading :
Prof. Mohinder Pal

Printed at : Ajanta Art Press / 2021/___

Syllabus and Course of Study in Statistics for B.A./B.Sc. (Semester IV)
for the Examinations to be held in the years 2017, 2018 and 2019

PAPER B: SAMPLING AND DESIGN OF EXPERIMENTS

UNIT – I

Complete Enumeration Vs Sample enumeration; advantages and disadvantages of sample survey, objectives of sampling, principle steps in sample survey, limitations of sampling, sampling and non sampling errors, types of sampling, probability sampling, purposive sampling and mixed sampling, random numbers. Simple Random sampling from finite population, S.R.S. with & without replacement, sample mean as unbiased estimate of population mean, sampling variance as an unbiased estimate of population mean, merits and demerits.

UNIT – II

Meaning of Stratification, Method of Stratified sampling and its advantages and disadvantages. Mean and variance of Stratified sampling, Proportional allocation, optimum allocation, comparison of Stratified random sampling with S.R.S.

UNIT – III

Systematic sampling, Cluster sampling with equal and unequal cluster sizes, estimation of mean and variance.

UNIT – IV

Analysis of variance for one way and two way classification, principles of design of experiment, randomization, replication and local control, concept and analysis of completely randomized design, randomized block design, advantages and disadvantages of these designs.

UNIT – V

Concept and analysis of Latin Square Design, one missing plot technique, Factorial experiment, their advantages, Factorial experiments for 2^2 and 2^3 designs, main effects and interactions.

SAMPLING THEORY

STRUCTURE

1.1 Objectives

1.2 Introduction

1.3 Definitions

1.4 Complete enumeration vs Sample enumeration

1.5 Advantages and Disadvantages of Sample survey

1.6 Objectives of Sampling

1.7 Principle steps in Sample Survey

1.8 Limitations of Sample

1.9 Sampling and Non Sampling errors

1.10 Self Assessment

1.11 Summary

1.1 OBJECTIVES

- To introduce Sampling Survey
- To introduce the advantages and disadvantages of sample survey
- To introduce principle steps in sampling survey

- To introduce errors in sampling
- To introduce self assessments based on sampling survey

1.2 INTRODUCTION

Sampling denotes the selection of a part of aggregate statistical material with a view to obtain information about the whole. " Sampling is not mere substitution of a partial coverage for a total coverage. Sampling is the science and art of controlling and measuring reliability of useful statistical information through the theory of probability". Before giving the idea of sampling it is essential to know about the population. In a statistical investigation the interest usually lies in the assessment of general magnitude and the study of variation with respect to one or more characteristics relating to individual belonging to a group. This group of individual under study is called **Population** or **Universe.**

The enumeration of population by sampling methods, proposed by Laplace in 1783, came into widespread use only by the mid-thirties of this Century. From the outset, some basic question arouse:

i) How should the observation be made?

ii) How many observations should be made?

iii) How should the total sample be made?

iv) How should the data thus obtained be analyzed?

The answers of these questions were sought and in the process, a number of different techniques and methods were tested to determine whether the above mentioned questions were adequately answered or not.

In other words "sampling is an inductive method of securing knowledge". We may have to secure knowledge about a large number of elements called **Population** or **Universe.** The only way to secure knowledge correctly is to examine every element separately. But it is not generally possible or practicable to study the entire population. So only few elements are chosen from it and the part consisting of chosen elements is called a sample and the method of selecting some elements is known as sampling method.

A sample should consist of two or more elements which would be representative of the characteristic of population. A sample throws light on the characteristic of the population. It must be noted that the sample should have more than one unit otherwise we estimate the error involved. In everyday life people make use of sampling to make judgments and take actions on the basis of information gathered from the study of the sample. A person tastes a grape or two before buying one or two kg.

1.3 DEFINITIONS

1.3.1 Population

In Statistics, the term **Population** is used in an altogether different sense its literary sense, "It is the totality of persons, objects, items or anything conceivable pertaining to certain characteristics". According to Kendial and Buckland (1975), the population is defined as, " in statistical usage, the term population is applied to any finite or infinite collection of individuals".

For example, the population of students in a University, number of plants in a field, persons suffering from cancer, workers in textile industry are some examples of population.

1.3.2 Finite Population

When the number of members of the population can be expressed as a definite quantity, the population is said to be **Finite.** For example, the population of books in the college library.

1.3.3 Infinite Population

A population is said to be infinite if it contain infinite number of elementary units or the composition is such that the elements of the population can't be counted. For example, hair on human body, number of stars in the sky etc.

1.3.4 Parameter and Statistic

Any population constant is called a parameter. For example, population mean (μ) and population variance (σ^2) etc. are parameters. Similarly, a statistic is a function of observable random variables and does not involve any unknown parameter. All the more, the function itself is a random variable. But a statistic is not necessarily an estimator of

some population parameter. For example $\frac{1}{n}\sum_{i=1}^{n} X_i$ i.e. $\bar{X}$ is a statistic.

1.3.5 Sampling Unit

The population may be regarded as consisting of units which are to be used for the purpose of sampling. Each unit is regarded as individual and such a unit is known as a sampling unit.

1.3.6 Sample size

The size of the sample is the no. of sampling units which are selected from a population by a random method. In brief, the number of sampling units selected in a sample is called sample size.

1.3.7 Sampling Frame

A complete list of sampling units which represents the population is called sampling frame.

1.3.8 Real Population

A population is said to be real or true if it contains existing objects with the factual observations. For example, the employees of J&K Government at certain date and time.

1.3.9 Standard Error

It is standard deviations of the sampling distribution of an estimator. The idea is that if we draw a number of repeated samples of fixed size "n" from a population having mean (μ) and variance (σ^2), each sample mean say $\bar{X}$, will have a different value. Here itself is a random variable and hence it has a distribution. The standard deviation of is called standard error. It has been proved that the standard error of the mean based on a sample of size "n"

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad \text{.....(1)}$$

From equation (1), it is obvious that larger the sample size smaller is the standard error and vice versa.

In practice we avoid studying or surveying the whole population. The process of drawing repeated samples is still more cumbersome. Hence, in reality neither we use "σ" to calculate the standard error of "$\bar{X}$" nor we take more than one sample. As a matter of fact, what we do is, that we select only one sample, find its standard deviation " s" and use the following formuls to find out the standard error of "$\bar{X}$" i.e.

$$S.E. = \frac{s}{\sqrt{n}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Standard error is commonly used in testing of hypothesis and interval estimation.

1.3.10 Hypothetical Population

A population is said to be hypothetical, if it is imaginary. For example, the possible outcomes of rolling a die "n" times.

1.4 COMPLETE ENUMERATION VS SAMPLE ENUMERATION OR CENSUS METHOD VS SAMPLE SURVEY METHOD

There are two ways of collection of statistical data

i) Census method or complete enumeration.

ii) Sample survey or sample enumeration.

In complete enumeration or census, information is collected from each and every member of the population, whereas sample survey refers to the study of the population on the basis of information obtained from sample. Census of human population in India and other countries of the world is an example of complete enumeration. When only a part of the population (called sample) is examined then we have a sample enumeration or sample survey e.g. estimating the crop yield of a country or counting the no. of trees in a forest.

In census method, since the entire population is investigated, it would require a large numbers of enumerators and other staff involving considerable time, money and labour. It is therefore generally applied for intensive investigations or when the number of members in the population is not large.

The sample survey method involving only a small part of the population has in comparison with the census method, greater flexibility as regards intensity of data collection and degree

of accuracy of final results, depending on the amount of time and money available for the purpose. Here sampling errors, no doubt influence the final conclusions, but non-sampling errors can be kept to a minimum because of deployment of highly trained personnel. A sample survey method is suitable for all situations, but in some cases census method cannot be applied.

1.5 ADVANTAGES AND DISADVANTAGES OF SAMPLE SURVEY

Sample survey method has defined advantages over census method from several stand points:

i) Sample survey takes less time, labour and money than is necessary in complete enumeration, because only a part of whole population is investigated. In most cases, time and money available for the completion of an investigation are limited and as such it becomes necessary to employ sample survey method.

ii) There is greater scope in sample survey than in census. Only a small group for skilled investigators is employed for collection of information. Although the cost per investigator is large in sample survey than in census, because of the specialized training to be imparted to these personnel, the amount of information collected by each is much larger, as such, a larger geographical area can be covered and more intensive data collection can be made in sample survey with the same amount of money spent in census.

iii) An extra advantage of sample survey method is that the magnitude of error is known. Sample survey depends on the laws of the probability, and such the magnitude of sampling errors can be theoretically calculated. Since non-sampling errors do not follow any definite probability law, no such estimate of error is possible in census method.

iv) In some cases, complete enumeration is not feasible and sample survey method is the only way. For example, a rice merchant cannot afford to examine every single grain of rice he purchases. He has to depend on a sample, based on which he forms an idea about the quality of rice in the whole consignment. Similarly to know the life hour of an electric bulb of some quality, we perform an experiment we get a sample of bulbs of same quality and observed the time (in hours of which they remain in working condition before they fuse). The mean life hours of these bulbs gives the average life period of the bulbs of the same quality. In this case we cannot experiment with all the bulbs of this quality, so the study of

a sample is the only way-out in this case. Also, in blood testing of a human being only few drops from the body of the person are examined. For such destructive testing the use of the sample is the only procedure of investigation.

Sample survey method has defined disadvantages over census method in the following way:

A great limitation of sample survey method is that it yields results for the concerned phenomenon as a whole or its broad subdivisions, but not for small sectors. Therefore, if time and cost are not important factors or the population is not very large, census method will be more appropriate than sample survey method.

1.6 OBJECTIVES OF SAMPLING

The main objectives of sampling are:

1. To obtain the maximum information about the population with the minimum effort.

2. To state the limits of accuracy of estimates based on samples.

3. Sampling is the process of learning about the population on the basis of a sample drawn from it. In sampling instead of studying each and every unit of the universe, a part of it is studied and the conclusions are drawn on that basis for the entire population.

4. Sampling methods are accepted for bringing efficiency in the estimation process.

5. The sample survey is made to obtain the specific information needed in the various fields such as social survey, economic survey, business survey etc.

1.7 PRINCIPLE STEPS IN A SAMPLE SURVEY

Sample survey techniques have now come to be used widely as an organized and established fact finding instrument and it is, therefore essential to describe briefly the main steps which are involved in a sample survey. Some of the main steps to be included are given as follows:

1.7.1 Specification of objective of survey

The aim and objectives of the survey must be very clearly explained to the experimenter so that unnecessary data collection may be avoided.

1.7.2 Definition of Population

The population from which the sample is to be selected must be clearly defined. For example, to estimate the average yield per plot for a crop, it is necessary to define the size of the plot in clear terms. The sampled population should coincide with the target population. The demographic, geographical, administrative and other boundaries of the population must be specified so that there remains no ambiguity regarding the coverage of the survey.

1.7.3 Choice of Sampling Units

The sampling units must together constitute the whole population and they must be distinct and non-overlapping. Also the complete list of sampling units must be made available.

1.7.4 Making of Questionnaire

The questionnaire requires a special aptitude which is attained by sharp intelligence and experience. The questionnaire should be small, simple non-overlapping, Questionnaire should be well planned and arranged in a logical sequence according to the objectives of the survey.

1.7.5 Selection of Proper Sample Design

Selection of an appropriate sampling technique provides more reliable results. Thus the sampling technique should be selected with care.

1.7.6 Organization of field work and collection of Data

The success of a sample survey completely depends upon the reliable field work. The field workers should correctly locate the selected units and collect information correctly according to the objectives of the survey.

1.7.7 Classification and Presentation of Data through Graphs, Charts

After collection of data the next step is to classify the collective data into tabular form on the basis of certain categories and then it can be presented through graphs and charts suich as Bar Diagram, Histogram, Pie Chart etc.

1.7.8 Analysis of Data

After classification of collected data a proper analysis is to made. Appropriate statistical technique or method must used to obtain the good results.

1.7.9 Report writing

The last step is to prepare the final report of the sample surveys as per the objectives.

1.8 LIMITATIONS OF SAMPLING

1. In spite of the fact that a proper choice of design is employed, a sample does not fully cover the parent population and consequently results are not exact.
2. Sampling theory and its applications in the field need the services of trained and qualified personnel without whom result of sample survey are not dependable.
3. The planning for execution of sampling survey should be done very carefully otherwise the data may provide misleading results.

1.9 NON-SAMPLING ERROR

An error in sample estimates cannot be attributed to sampling fluctuations only. It is expected that the studies based on complete enumeration do not yield similar result if repeated enumeration is done. Such a discrepancy occurs due to many errors which are termed as non sampling errors. In brief, the main sources of non sampling errors are:

a) Failure to measure some of the units in the selected sample.

b) Observational errors due to defective measurement.

c) Errors introduced in editing, coding and tabulating the results.

In practice the census survey results may suffer from non sampling errors although these may be free from sampling errors. The non sampling errors are likely to increase in sample size, while sampling error decreases with increase in sample size.

1.10 SELF ASSESSMENT

Q.1. Write a note on sampling theory.

Q.2. Define with examples.

i) Population, Infinite and finite population.

ii) Real Population.

iii) Sampling unit.

iv) Standard error.

Q.3. Write a short note on sampling and non sampling errors.

Q.4. Difference between Census and Sample survey method.

Q.5. What are the various advantages and disadvantages of sample survey.

1.11 SUMMARY

The total count of all units of the population for a certain characteristic is known as complete enumeration, also termed census survey and when only a part, called a sample, is selected from the population and examined, it is called sample enumeration or sample survey. Despite the above advantages sample survey is not always preferred to census surveys. Sampling theory has its own limitations and advantages over complete enumeration.

In the sample survey, the final step of the analysis and drawing inferences from a sample to a population is a very vital and fascinating issue. Since the results of the survey are the basis for the policy making, it is the most essential part of the sample survey and should be handled carefully**.**

TYPES OF SAMPLING

STRUCTURE

2.1 Objectives

2.2 Introduction

2.3 Basic Principles of sampling

3.4 Types of sampling

2.5 Random Numbers

2.6 Illustrations

2.7 Summary

2.1 OBJECTIVES

- To introduce the types of sampling
- To introduce random sampling
- To introduce some problems and self assignments based on sampling

2.2 INTRODUCTION

The technique of selecting a sample is of fundamental importance in sampling theory and usually depends upon the nature of the investigation. The sampling methods are mainly used for opinion surveys, but cannot be recommended for general use as it is subject to drawbacks of prejudice and bias of the sampler. However, if the sampler is experienced and an expert, it is possible that judgment sampling may yield useful results. It however,

suffers a serious defect that it is not possible to compute the degree of precision of the estimate from the sample values.

In the early days of the data analysis, people were accustomed to deal with the complete enumeration as they were suspicious about the sampling results. There suspicion might have arisen because in the past sampling techniques were carelessly handled and they were rather misused during the survey work. But now, things are very much changed in the reference. The sampling techniques have sufficiently been advanced and surveys have become cautions and alert against their possible errors and misuses. Subsequently these methods, now a day, are quite confidently used all around. In fact, sampling these days is a valuable tool for obtaining the data quickly, accurately and above all, economically. Not only this, in many situations the complete enumeration has become impossible and the use of sampling technique have as such become unavoidable.

2.3 BASIC PRINCIPLES OF SAMPLING

The theory of sampling is based on the following principles which are known as the basic principles of sampling:

i) The principle of Statistical Regularity. ii. Principle of optimization

iii) The principle of inertia of large numbers. iv. Principle of validity

The law of Statistical Regularity according to King is " a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group".

The law of inertia of large numbers states that, "other things being equal, as the sample size increases, the result of the sample tends to be more reliable and accurate i.e. goes near to the population values". These important laws determine the possibility of reaching valid and optimum conclusions about a population on the basis of sample. By validity we mean that a sample design enables us to obtain valid tests and estimates. By optimization we mean an increase in efficiency and decrease in cost.

2.4 TYPES OF SAMPLING

The process of selecting a sample is often called sampling methods or sampling technique. It distinguishes one type of sample from another. The sampling techniques mainly depend

upon the purpose of survey. There are three main techniques of selecting a sample:

i) Subjective or Purposive sampling

ii) Probability or objective sampling

iii) Mixed sampling

According to these sampling procedures the samples may also be classified into three classes:

i) Subjective sample

ii) Probability sample

iii) Mixed sample

SUBJECTIVE OR PURPOSIVE SAMPLING:-

This is the easiest technique of sampling in which one selects a desired number of sampling units which he thinks representative of the population according to his own criterion. Here, only those units are selected which are convenient to the investigator and he takes them relevant to the purpose in view is likely to suffer from the drawback of favoritism and personal bias and as such may not remain a true representative of the population.

For instances, extremists, politicians, organizations, advertisers, labour units etc. offer information only from those people who are expected to respond as per their desired motives. They use or misuse these results taking representative of the entire population.

This type of sampling remains certainly biased. They may present satisfactory results when the sample size is small. They remain very much biased and as such are quite misleading when the sample size is large. Since, no theoretical or mathematical approach is possible to this method. Hence, it is used very rarely and only in extreme cases.

PROBABILITY OR OBJECTIVE SAMPLING:-

In this technique the samples are selected according to some laws of chance (probability) in such a way that each unit of the population has a known or definite probability of being selected in the sample.

In this type of sampling:

i) Every sampling unit may be assigned an equal probability of being selected.

ii) Different sampling units may have different probabilities of being selected.

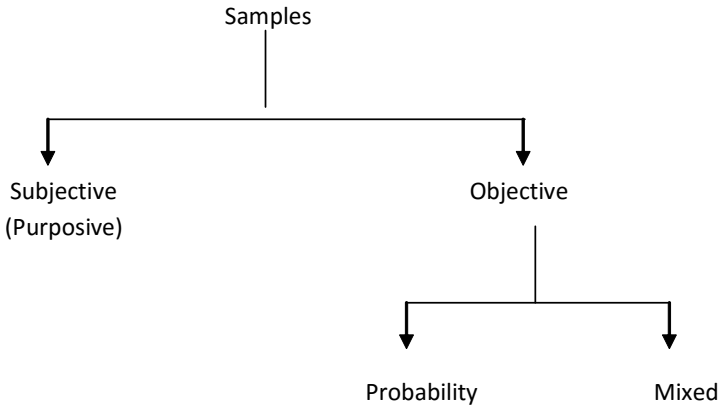
iii) A sampling unit may be assigned the probability of selection in proportional to its size.

The samples are drawn according to some probability law are called probability samples. In brief, probability sampling is the method of selecting samples according to certain laws of probability in which each unit of the population has some definite probability of being selected in the sample. It is to be noted here that there are numbers of samples specified type S_1, S_2 --- S_k that can be formed by grouping units of a given population and each possible sample S_i has, assigned to it, a known probability of selection P_i. A clear specification of all possible samples of a given type along with their corresponding probabilities of selection is said to constitute a sample design. In subsequent chapters of this study material, we shall consider only the procedures of probability sampling.

MIXED SAMPLING:

The technique in which samples are selected partly by some probability rule and partly like a purposive sampling is called mixed sampling. The samples drawn by this method are called mixed samples.

Remarks: The samples may also be classified as shown below:



These sampling approaches have given a number of the sampling methods, some of them are given below:

1. SRSWOR
2. SRSWR
3. Stratified Random Sampling
4. Systematic Sampling
5. Multistage Sampling
6. Quasi Random Sampling
7. Double Sampling
8. Area Sampling
9. Simple Cluster Sampling
10. Multistage Cluster Sampling
11. Quota Sampling

2.5 RANDOM NUMBERS

Random numbers refers to some well known sequence of digit in which the successive figures are in a perfectly random order. This means that if a digit is blindly selected from random number table, any of the ten digits 0, 1, 2, 3, 4, 5 - - - 9 is likely to occur with the same probability 1/10. Similarly, if two consecutive digits are taken, the number formed by them (in the order they appear) may be any of the hundred number 00, 01, 02, 03, - - - 99 with the same probability 1/100 and so on. The series of random numbers are usually available in groups of four digits to facilitate easy reading as follows:

5404	9235	3822	3756	0653
1014	7894	9307	0458	9983
7515	4258	9200	5866	7302

Of course, any haphazard arrangement of the digits 0 to 9 will not give random number. The random number have been prepared by special devices and tested for randomness. Some of the famous series of random number are those of Tippett (41,600 digits); Kendall and Bibington Smith (100,000 digits); Fisher and Yates (15,000 digits); Rand Corporation (1,000,000) random numbers are used for drawing random samples. In brief, random numbers are the digits 0, 2, 3, 4, 5, 6, 7, 8, 9 scrambled in a particular fashion or in the

form of a particular series. Most of these series are the result of the actual sampling operations recorded for the future use.

(i) Tippett Series: One such set is given by L.H.C. Tippett consisting of 41,600 digits combined in fours to give 10400 four digit numbers. Though Tippett constructed the series by choosing digits from census in a haphazard fashion, yet their application in the various investigations has proved them true and quite dependable.

(ii) Fisher and Yates Series: Fisher and Yates table consists of 15,000 numbers arranged in two digits in 300 blocks. These numbers were picked up from the 10^{th} to 19^{th} digits of A.S. Thompson's 20 – figure logarithmic tables and were subsequently adjusted through the results of playing cards just to introduce an element of randomness in the selection of digits and in allotting the selected digits in to the blocks (because it was found there were too many six in the parent logarithmic table).

(iii) Kendall and Babington Smith's number: This table of random numbers consists of 1,00,000 numbers grouped in two and four digits and are arranged in 100 separate blocks of 1000 each. These numbers were obtained with the help ofspecially refined gambling machine of toothed wheel. Besides these some more series of random numbers are also available.

Here an obvious questions arises that what is the guarantee of randomness of these numbers? Definitely there is no theoretical argument in support of it, but guarantee lies in practical tests. All these random numbers series have been subjected successfully to a large number of investigations and examinations. It is experienced that in practice these series are highly reliable.

2.6 ILLUSTRATIONS

i) Suppose a random number of 5 students is to be taken out of a college having the strength of 1236. Making all the students serially numbered from 0001 to 1236, any page of random number series is taken and starting from any point of it either row wise or column wise four digit numbers are noted (as 1236 consists 4 digits) ignoring the values greater than 1236.

The following noted 25 random numbers are given by Kendall and Smith in the series of

first thousands taken column wise in the digit of 4

2315	1174	0709	0924	9783
0554	4336	4331	9795	8999
1487	9380	6157	9373	2596
3897	4954	3135	7262	8184
9731	3676	5704	6102	1132

Thus the students bearing the numbers 0554, 1174, 0709, 0924 and 1132 will constitute our demand desired random sample of size 5.

2. Draw a random sample without replacement of size 7 from a population of size 836.

Ans. Mark the population units with numbers 001 to 836, then select a three digit number anywhere in the random number table an move row wise or column wise to select the three digit numbers, discarding the numbers and above 836 and we go on until the 7 numbers which are less than 836 are obtained. As such, the sample obtained will consist of the desired units.

Following are the 25 random sampling numbers from Fisher odd Yates table:

04	28	50	13	92
31	64	94	20	96
86	28	36	82	58
79	24	68	66	86
45	3	42	65	29

Starting from 042, the units bearing numbers 042, 316, 792, 451, 494, 836, 468 will form our random sample of size 7 without replacement. Here we have discarded 862 and 850 as they are greater than 836.

2.7 SUMMARY

The technique of selecting a sample is of fundamental importance in sampling theory and

usually depends upon the nature of the investigation. The sampling procedures which are commonly used may be broadly classified under the following heads:

(i) Purposive Sampling

(ii) Probability Sampling

(iii) Mixed Sampling

The sampling methods are mainly used for opinion surveys but cannot be recommended for general use as these are subject to the drawbacks of prejudice and bias of the samples.

SAMPLING THEORY
SIMPLE RANDOM SAMPLING

STRUCTURE

3.1 Objectives

3.2 Introduction

3.3 Definition of Simple Random Sampling (SRS)

3.4 Simple Random Sampling with Replacement (SRSWR)

3.5 Simple Random Sampling without Replacement (SRSWOR)

3.6 Illustrations

3.7 Self Assessment

3.8 Summary

3.1 OBJECTIVES

The objective of the sampling is to obtain maximum possible information with reliability about the population at the minimum possible cost, time and energy. The selection of sampling units in simple random sampling can be made by simple random sampling with replacement and without replacement.

3.2 INTRODUCTION

The simplest and most common method of sampling is simple random sampling (SRS) in which the sample is drawn unit by unit with equal probability of selection for each unit at

each draw. Therefore simple random sampling (SRS) is a method of selecting "n" units out of a population of size "N" by giving equal probability to all the units or a sampling procedure in which all the possible combinations of "n" units that may be formed from the population of "N" units have same probability of selection. It is also sometimes referred to as unrestricted random sampling. If a unit is selected and noted and then returned to the population before the next draw is made and this procedure is repeated "n" times, it give rise to a simple random sampling with replacement (SRSWR). If this procedure is repeated till "n" distinct units are selected and all the repetitions are ignored, it is called simple random sampling without replacement (SRSWOR)

3.3 DEFINITION OF SIMPLE RANDOM SAMPLING

A random sample is said to be simple if each and every unit of the population has an equal probability of being selected in the sample. Thus simple random sampling is that particular case of probability sampling in which each unit of the population has an equal and independent chance of being included in the sample.

In brief, if a unit is selected and noted and then returned to the population before the next draw is made and this procedure is repeated n times, it give rise to a simple random sampling of "n" units and this procedure id generally known as simple random sampling.

3.4 SIMPLE RANDOM SAMPLING WITH REPLACEMENT (SRSWR)

Simple random sampling is said to be "with replacement" when the sample members are drawn from the population one by one and after each draw, the selected sample unit is noted and returned to the population before the next one is drawn. This means that at each stage of the sampling process all the population units (including those obtained in earlier draw) are considered for the selection with equal probability. Thus the population remains the same before each draw and any of the population units may appear more than once in the sample.

In brief, if the sampling unit drawn is replaced back in the population before next draw is the case of SRSWR and the units are drawn from the whole of the population with equal probability.

3.5 SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT (SRSWOR)

Simple random sampling is said to be "without replacement" when either the sample members are drawn all at a time or drawn one by one in such a manner that after each drawing the selected unit is not returned to the population when the next one is drawn. This means that when drawing is made one by one at each stage of the sampling process the population units already chosen are not considered for subsequent selections but the drawing is made with equal probability only from those units not selecting in any of the earlier drawings. It is evident that in simple random sampling without replacement (SRSWOR) from a finite population, the size of the population goes on diminishing as the sampling process continues. Consequently, no population unit can appear more than once in the sample.

In brief, if the unit is not drawn is not replaced then we call it simple random sampling without replacement with "N" units in the population, the probability of selection of any unit at first draw is $\frac{1}{N}$, at second draw is $\frac{1}{N-1}$ and so on. It means in SRSWOR, any unit cannot appear more than once in the sample

3.6 ILLUSTRATIONS

Problem: Derive the bias and mean square error of $\frac{N}{n}\sum_{i=1}^{n} y_i$ under probability proportional to size sampling with replacement.

Solution: Bias of the estimator is given by

$$B = \frac{N}{n}\sum_{i=1}^{n} E(y_i) - y$$

$$B = \frac{N}{n}\sum_{i=1}^{n} \sum_{j=1}^{n} y_i P_i - N\bar{y}$$

$$B = \frac{N}{n} n \sum_{j=1}^{n} (y_i - \bar{y}) P_i$$

$$B = N \sum_{j=1}^{n} (y_i - \bar{y}) P_i$$

consider the difference

$$\frac{N}{n}\sum_{i=1}^{n} E(y_i) - x = \frac{N}{n}\sum_{i=1}^{n}(y_i - \bar{x})$$

Squaring both side and taking expectation on both sides, we get the mean square error as

$$H = \frac{N^2}{n^2}\left[\sum_{i=1}^{n} E(y_i - \bar{y})^2 + 2\sum_{i=1}^{n} E(y_i - \bar{y})\ y_i - \bar{y}\right]$$

$$= \frac{N^2}{n^2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\ (y_i - \bar{y})^2\ P_j + \sum_{i<j} 0\ \right]$$

$$= \frac{N^2}{n^2}\left[\sum_{i=1}^{n}(y_i - \bar{x})^2\ P_i\right]$$

Cross product terms becomes zero because units are drawn independently one by one with replacement.

Problem: In a population of 4000 people who were called for casting their votes, 50 percent returned to the polls. Estimate the sample size to estimate this proportion so that the marginal error is 5 % with 95 percent confidence coefficient when the sampling is done

i) With replacement

ii) Without replacement

Solution:

i) Sampling with replacement: In this case, we can ignore f_{pc} and have

$$n_0 = \frac{t^2 p(1-p)}{\epsilon^2}$$

$$= \frac{(1.96)^2(0.5)(0.5)}{0.0025}$$

$$= 384.16$$

(ii) Sampling without replacement:

$$n_1 = \frac{n_0}{1+\frac{(n_0-1)}{N}} = 352$$

3.7 SELFASSESSMENT

Q.1. Define Simple Random sampling.

Q.2. How many types of Simple random sampling. Name them write briefly.

Q.3. Write the properties of SRS.

Q.4. Write the properties of SRSWR and SRSWOR.

Q.5. Give some examples on SRS.

3.8 SUMMARY

Random sampling is the simple and most important among the various sampling techniques. It is absolutely free from the influence of human bias, and hence also called lottery sampling. Random sampling is the most appropriate in case when the population is more or less homogeneous with respect to the characteristic under study. Similarly, simple random sampling is a special case of SRS in which the probability of selection of any particular member remains constant throughout the sampling process irrespective of whether the member had been selected earlier or not. Therefore SRSWR will always give a simple sample from a finite or infinite population. However, SRSWOR gives a simple sample only when the population is finite.

MEAN AND VARIANCE OF SIMPLE RANDOM SAMPLING

STRUCTURE

4.1 Objectives

4.2 Introduction

4.3 Mean and variance of Simple random sampling

4.4 Merits and Demerits of SRS

4.5 Illustrations

4.6 Summary

4.1 OBJECTIVES

- To introduce Simple Random Sampling
- To introduce mean and variance of Simple Random Sampling

4.2 INTRODUCTION

A random sample is said to be simple if each and every unit of the population has an equal probability of being selected in the sample. Thus simple random sampling (SRS) is that particular case of probability sampling in which each unit of the population has an equal and independent chance of being included in the sample.

The simplest method of probability sampling is SRSWOR. This method is based on selecting the sample units one by one by assigning equal probability of selection at the first and each subsequent draw. Notations of SRS for mean and variance are:

Population total: $Y = Y_1 + Y_2 + Y_3 + ---+ Y_N$

$= \sum_{i=1}^{N} Y_i$

Population Mean: $\bar{Y} = \frac{Y_1+Y_2+Y_3+---+Y_N}{N}$

$=> \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

Sample total: $y = y_1 + y_2 + --- + y_n$

$= \sum_{i=1}^{n} y_i$

Sample mean: $\bar{y} = \frac{y_1+y_2+---+y_n}{n}$

$=> \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Population Variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ if the population is infinite

$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ if the population is finite

Sample Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

4.3 MEAN AND VARIANCE

Theorem 1: In SRSWOR the sample mean $\bar{y}$ is an unbiased estimate of the population mean $\bar{Y}$ or

Prove that $E(\bar{y}) = \bar{Y}$

Proof: Since $E(\bar{y}) = E[\frac{1}{n}\sum_{i=1}^{n} y_i]$

$= \frac{1}{n}\sum_{i=1}^{n} E(y_i)$

$= \frac{1}{n}\sum_{i=1}^{n}[\frac{Y_1}{N} + \frac{Y_2}{N} + --- + \frac{Y_N}{N}]$

[because $E(y_i)=\frac{1}{N}\sum_{i=1}^{n} Y_i$]

$$=\frac{1}{n}\sum_{i=1}^{n}\ \left[\sum_{i=1}^{n}\frac{Y_i}{N}\right]$$

$$=\frac{1}{n}n\ \frac{1}{N}\sum_{i=1}^{n} Y_i = \bar{Y}$$

Hence, sample mean is an unbiased estimate of population mean.

Theorem 2: In SRSWOR the variance of the sample mean $\bar{y}$ is given by

$$V(\bar{y}) = \frac{S^2}{n}\frac{N-n}{N} = (\frac{1}{n}-\frac{1}{N})S^2$$

Where $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$

Proof: We know that

$$V(\bar{y}) = E[\bar{y} - E(\bar{y})]^2$$

$$= E[\bar{y} - \bar{Y}]^2$$

$$= E[\frac{1}{n}\sum_{i=1}^{n} y_i - \bar{Y}]^2$$

$$= \frac{1}{n^2}\ E[\sum_{i=1}^{n} y_i - n\bar{Y}]^2$$

$$= \frac{1}{n^2}\ E[y_1+y_2 + - - - -+y_n - n\bar{Y}]^2$$

$$= \frac{1}{n^2}\ E[(y_1 - \bar{Y})\ + (y_2 - \bar{Y})\ +\ - - -\ + (y_n - \bar{Y})\ ]^2$$

$$= \frac{1}{n^2}\ E[\sum_{i=1}^{n}(y_1 - \bar{Y})^2 + \sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - \bar{Y})(y_j - \bar{Y})]$$

$$= \frac{1}{n^2}\ [\sum_{i=1}^{n} E(y_i - \bar{Y})^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} E(y_i - \bar{Y})(y_j - \bar{Y})]$$

where $(i \neq j)$

$$\Rightarrow V(\bar{y}) = \frac{1}{n^2}\left[n\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2 + \frac{n(n-1)}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - \bar{Y})(y_j - \bar{Y})\right]$$

where $(i \neq j)$

$$= \frac{1}{n^2}\left[\frac{n}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2 - \frac{n(n-1)}{N(N-1)}\sum_{i=1}^{N}(Y_i - \bar{Y})^2\right]$$

because $\left[\sum_{i=1}^{N}\sum_{j=1}^{N}E(Y_i - \bar{Y})(Y_j - \bar{Y}) = -\sum_{i=1}^{N}(Y_i - \bar{Y})^2\right]$

where $(i \neq j)$

$$= \frac{1}{n^2}\left[\frac{n}{N}(N-1)S^2 - \frac{n(n-1)}{N(N-1)}(N-1)S^2\right]$$

$$= \frac{1}{n^2}\left[\frac{n}{N}(N-1)S^2 - \frac{n(n-1)}{N}S^2\right]$$

$$= \frac{S^2}{Nn}[N-1-n+1]$$

$$= \frac{S^2}{Nn}[N-n]$$

$$= \left[\frac{N}{nN} - \frac{n}{Nn}\right]S^2$$

$$\Rightarrow V(\bar{y}) = \left[\frac{1}{n} - \frac{1}{N}\right]S^2$$

Hence the theorem is proved

Theorem 3: In SRSWOR, the sample variance s^2 is an unbiased estimate of the population variance S^2, where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

and $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$

Proof: We know that

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Add and Subtract $\bar{Y}$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y} + \bar{Y} - \bar{Y})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}[(y_i - \bar{Y}) - (\bar{y} - \bar{Y})]^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}[(y_i - \bar{Y})^2 + (\bar{y} - \bar{Y})^2 - 2(y_i - \bar{Y})\,(\bar{y} - \bar{Y})\,]$$

$$= \frac{1}{n-1}[\sum_{i-1}^{n}(y_i - \bar{Y})^2 + n(\bar{y} - \bar{Y})^2 - 2\ (\bar{y} - \bar{y})\sum_{i=1}^{n}(y_i - \bar{Y})\,]$$

$$= \frac{1}{n-1}[\sum_{i-1}^{n}(y_i - \bar{Y})^2 + n(\bar{y} - \bar{Y})^2 - 2\ (\bar{y} - \bar{Y})\,(n\bar{y} - n\bar{Y})]$$

because $\sum_{i=1}^{n}(y_i - \bar{y}) = \frac{n}{n}\sum_{i=1}^{n} y_i - n\bar{Y} = (n\bar{y} - n\bar{Y})$

$$= \frac{1}{n-1}[\sum_{i-1}^{n}(y_i - \bar{Y})^2 + n(\bar{y} - \bar{Y})^2 - 2\ n(\bar{y} - \bar{Y})^2\,]$$

$$\Longrightarrow s^2 = \frac{1}{n-1}[\sum_{i-1}^{n}(y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2\,]$$

Take Expectation both sides, we get

$$E(s^2) = \frac{1}{n-1}[\sum_{i-1}^{n} E(y_i - \bar{Y})^2 - nE(\bar{y} - \bar{Y})^2\,]$$

$$= \frac{1}{n-1}[\frac{n}{N}\sum_{i-1}^{n} E(Y_i - \bar{Y})^2 - nV(\bar{y})\,]$$

$$= \frac{n}{n-1}[\frac{1}{N}\sum_{i-1}^{n} E(Y_i - \bar{Y})^2 - V(\bar{y})\,]$$

$$= \frac{n}{n-1}[\frac{N-1}{N}S^2 - \frac{S^2}{n}\frac{N-n}{N}]$$

$$\Rightarrow E(s^2) = \frac{n}{n-1}\frac{S^2}{N}\left[\frac{Nn-n-N+n}{n}\right]$$

$$= \frac{n}{n-1}\frac{S^2}{N}\frac{N(n-1)}{n}]$$

$$\Rightarrow E(s^2) = S^2$$

Theorem 4: Show that in SRSWOR, the sample variance is less than the variance in case of SRSWR.

or

Prove that SSWOR is more efficient than SRSWR.

or

Prove that the variance of sample mean is more in the SRSWR as compare to its variance in SRSWOR.

Proof: Before the proof of this theorem we first find $V(\bar{y})$ in SRSWR as

$$V(\bar{y}) = E[\bar{y} - E(\bar{y})]^2$$

$$= E[\bar{y} - \bar{Y}]^2$$

$$= E[\frac{1}{n}\sum_{i=1}^{n} y_i - \bar{Y})^2$$

$$E[\frac{y_1+y_2+---+y_n}{n} - \bar{Y}]^2$$

$$= \frac{1}{n^2}E[y_1 + y_2 + --- + y_n - n\bar{Y}]^2$$

$$= \frac{1}{n^2}E[(y_1 - \bar{Y}) + (y_2 - \bar{Y}) + --- + (y_n - \bar{Y})]^2$$

$$\Rightarrow V(\bar{y}) = \frac{1}{n^2}E[\sum_{i=1}^{n} (y_i - \bar{Y})^2 + \sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - \bar{Y})(y_j - \bar{Y})]$$

where $(i \neq j)$

$$\Rightarrow V(\bar{y}) = \frac{1}{n^2}[\sum_{i=1}^{n} E(y_i - \bar{Y})^2 + \sum_{i=1}^{N}\sum_{j=1}^{N} E(y_i - \bar{Y})(y_j - \bar{Y})]$$

where $(i \neq j)$

$$= \frac{1}{n^2}[nV(y) + n(n-1)E(y_i - \bar{Y})E(y_j - \bar{Y})]$$

$$= \frac{1}{n^2}[nV(y) + 0.0]$$

[Because sum of the deviations taken from A.M. is always equal to zero]

$$\Rightarrow V(\bar{y}) = \frac{1}{n^2} nV(y)] = \frac{V(y)}{n}$$

$$\Rightarrow V(\bar{y})_{SRSWR} = \frac{\sigma^2}{n} \text{ ———(i)}$$

We know that in case of SRSWOR

$$V(\bar{y})_{SRSWOR} = \frac{N-n}{N}\frac{S^2}{n}$$

$$= \frac{N-n}{N}\frac{1}{n}S^2\frac{N-1}{N-1}$$

$$V(\bar{y})_{SRSWOR} = \frac{N-n}{N}\frac{1}{n}\frac{N\sigma^2}{N-1}$$

$$\Rightarrow V(\bar{y})_{SRSWOR} = \frac{\sigma^2}{N}\frac{N-n}{N-1} \text{ ——— (ii)}$$

Now compare $V(\bar{y})_{SRSWR}$ & $V(\bar{y})_{SRSWOR}$ i.e.

$$V(\bar{y})_{SRSWOR} = \frac{N-n}{N}\frac{\sigma^2}{n}$$

and

$$V(\bar{y})_{SRSWR} = \frac{N-n}{N}$$

Here $\frac{\sigma^2}{N}\frac{N-n}{N-1} < \frac{\sigma^2}{n}$

Hence $V(\bar{y})_{SRSWOR} < V(\bar{y})_{SRSWR}$ i.e. we can say that the variance of SRSWOR is more efficient than that of Variance of SRSWR.

Note : Since

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

$$\Rightarrow S^2(N-1) = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

$$\Rightarrow S^2(N-1) = \frac{N}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

$$\Rightarrow S^2(N-1) = N\sigma^2$$

4.4 MERITS AND DEMERITS OF SIMPLE RANDOM SAMPLING (SRS)

4.4.1 Merits of Simple Random Sampling

i. In SRS each and every unit of the population has an equal probability of being selected in the sample.

ii. In case of SRS, the sample mean is an unbiased estimate of the population mean.

iii. The selection technique of sampling units from the population is very simple.

iv. The cost of the survey in SRS is also less.

4.4.2 Demerits of SRS

i. If the population is heterogeneous then SRS provides less efficient estimator.

ii. In case of SRSwr any unit can appear more than once in the sample.

iii. The sample size in SRSWR may be more than the population size.

iv. Sometimes we may not able to get true representative of the population through simple random sample.

4.5 ILLUSTRATIONS

From a SRS of size n drawn from N unit's simple sub-sample of n_1 unit is duplicated and added to the original sample. Show that the mean based on $(n+n_1)$ units is an unbiased estimate of the population mean. Obtain the variance of the estimate. How does it compare

with the variance of the estimate based on n units only?

Solution: Let $Y_1, Y_2 - - - Y_n$ represents the values of the variable in the N units of the population and let $y_1, y_2, - - - y_n$ be a random sample of size n from this population. A sub-sample of size n_1 is drawn from the values $y_1, y_2, - - - y_n$ without any loss of generality we can assume the sub sample to consist of the observations $y_1, y_2, - - - y_n$. Thus the combined sample of size $n+n_1$ can be split into the following two samples.

i) A sample of size 2n, consisting of the values $y_1, y_2, - - - y_n$ each repeated twice, and

ii) A sample of size $n-n_1$ consisting of the values $y_{n_1+1}, y_{n_1+2} - - - y_{n_1 n}$.

The estimate $\bar{y}$ of the population mean $\bar{Y}$ base on the sample of size $n+n_1$ is given by

$$\bar{y} = \frac{2\sum_{i=1}^{n_1} y_i + \sum_{j=n_1+1}^{n} y_j}{n+n_1} = \frac{2n_1\bar{y}_1 + (n-n_1)\bar{y}_2}{n+n_1} \qquad \text{——————} (1)$$

Where $\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$ and $\bar{y}_2 = \frac{1}{(n-n_1)}\sum_{j=n_1+1}^{n} y_i$

Since sample mean is an unbiased estimate of the population mean, we get

$E(\bar{y}_1) = \bar{Y}_N$ and $E(\bar{y}_2) = \bar{Y}_N$

Taking expectations both sides on equation (1) we get

$$E(\bar{Y}) = \frac{1}{n+n_1}[2n_1 E(\bar{y}_1) + (n-n_1)E(\bar{y}_2)]$$

$$= \frac{1}{n+n_1}[2n_1\bar{Y}_N + (n-n_1)\bar{Y}_N]$$

$$\Longrightarrow E(\bar{Y}) = \bar{Y}_N$$

Hence $\bar{y}$ is an unbiased estimate of $\bar{Y}_N$ variance of the estimate. If we write

$U = \sum_{i=1}^{n_1} y_i = n_1\bar{y}_1$ and

$U = \sum_{j=n_1+1}^{n} y_i = (n-n_1)\bar{y}_2$

Then

$$\bar{y} = \frac{1}{n+n_1}[2U + V]$$

$$V(\bar{y}) = \frac{1}{(n-n_1)2}[4\text{Var}(u) + V(v) + 4\text{Cov}(u,v)] \qquad \text{——— (2)}$$

In SRSWOR, we have

$$\text{Var}(u) = n_1^2;$$

$$V(\bar{y}_1) = \frac{n_1(N-n_1)}{N} S_2$$

Also $\text{Cov}(u,v) = E[(u-\bar{u})(v-\bar{v})]$

$$= E[\sum_{i=1}^{n_1}(y_i - \bar{Y}_N)\sum_{j=n_1+1}^{n}(y_i - \bar{Y}_N)]$$

$$= \sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n} E(y_i - \bar{Y}_N)(y_j - \bar{Y}_N)$$

$$= \sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n} \text{Cov}(y_i, y_j)$$

$$= \sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}(-\frac{S^2}{N})$$

$$\Rightarrow \quad \text{Cov}(u,v) = -\frac{n_1(n-n_1)}{N} S^2$$

Where S^2 is the population mean square substitute these values in equation (2) we get

$$V(\bar{y}) = \frac{S^2}{N(n-n_1)^2}[Nn\left(1 + \frac{3n_1}{n}\right) - (n+n_1)^2]$$

Or original sample of size n, we have

$$V(\bar{y}_n) = \frac{N-n}{Nn} S^2$$

$$\frac{V(\bar{y})}{V(\bar{y}_n)} = \frac{n}{(N-n)(n+n_1)^2}[Nn\left(1 + \frac{3n_1}{n}\right) - (n+n_1)^2]$$

$$= \frac{\left(\frac{n}{n+n_2}\right)^2\left(1+\frac{sn_2}{n}\right)}{\left(1-\frac{n}{N}\right)} - \frac{\frac{n}{N}}{\left(1-\frac{n}{N}\right)}$$

If N is large as compared to n such that $\frac{n}{N}$ is ignored, then

$$\frac{V(\bar{y})}{V(\bar{y}_n)} = \frac{\left(1+\frac{sn_2}{n}\right)}{\left(1-\frac{n_2}{n}\right)^2}$$

Thus the relative loss in efficiency resulting from duplicating of a subset of elements is given by

$$\frac{V(\bar{y})-V(\bar{y}_n)}{V(\bar{y}_n)} = \frac{V(\bar{y})}{V(\bar{y}_n)} - 1$$

$$= \frac{\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)}{\left(1+\frac{n_1}{n}\right)^2}$$

In particular, if we take $\frac{n_2}{n} = \frac{1}{3}$ then,

Loss efficiency $= \frac{\frac{1}{3}\left(\frac{2}{3}\right)}{\frac{16}{9}}$ 0.125

4.6 SUMMARY

On the basis of the above, it can be concluded that the various steps in the process of sampling are interwoven. Moreover, the sample should be selected randomly and its size should be adequate enough to represent the universe.

STRATIFICATION

STRUCTURE

5.1 Objectives

5.2 Introduction

5.3 Meaning of Stratification

5.4 Methods of Stratified sampling

5.5 Advantages of Stratified sampling

5.6 Disadvantages of Stratified sampling

5.7 Mean and Variance OF Stratified sampling

5.8 Self Assessment

5.9 Summary

5.1 OBJECTIVES

- To introduce the meaning of Stratification.
- To introduce the method of Stratified sampling.
- Advantages and Disadvantages of Stratified sampling.
- Mean and Variance of Stratified Sampling.

5.2 INTRODUCTION

In simple random sampling, it has been seen that the precision of the standard error of the estimator of the population total (mean) depends on two aspects, namely the sample size and the variability of the character under study. Therefore, in order to get an estimator with increased precision one can increase the sample size. However, consideration of cost, limits the size of the sample. The other possible way to estimate the population total with greater precision is to divide the population into several groups each of which is more homogeneous than the entire population and draw sample of predetermined size from each of these groups. The groups into which the population is divided are called Strata and drawing sample from each of the strata is called Stratified sampling. In stratified sampling, samples are drawn independently from different strata and it is not necessary to use the same sampling designing in different strata. For example, in the absence of suitable size information, SRS can be used in few strata, where as probability proportional to size sampling can be used in the remaining strata when size information is available for those strata.

5.3 MEANING OF STRATIFICATION

We know that the precision (result or accuracy) of a sample estimate of the population mean depends not only upon the size of the sample and the sample fraction but also on the variability or heterogeneity of the population apart from the size of the sample. Therefore the only way to increase the precision of an estimate it is desirable to reduce the heterogeneity. One such procedure is Stratified random Sampling.

In Stratified Sampling N units are first divided into sub-populations or strata of sizes $N_1, N_2 \dots N_k$.The sub population is non overlapping so that $N_1 + N_2 + \dots + N_k = N$ i.e. sum of all the data is equal to N. The sub-population is called strata. When a strata has been determined a sample is drawn from each, the drawing is made independently in different strata. The sample size within the strata are $n_1, n_2 \dots n_k$ respectively. If a SRS is drawn from each stratum, the whole procedure is described as **Stratified** Random Sampling.

In brief, the homogeneous groups into which the population is divided are called

Strata and drawing sample from each of the strata is called Stratified sampling. This process is called stratification.

5.4 METHODS OF STRATIFIED SAMPLING

There is always an effort to adopt a sampling procedure which gives estimates with greater precision. Precision of a sample estimate depends on two factors these are:

i) The size n of the sample.

ii) The variability or heterogeneity of the population unit.

Because of the cost and time considerations the size n of the sample cannot be increased beyond a unit. Once the size of the sample is decided, the only way to increase the precision of the estimates is to reduce the heterogeneity of the population to the greater possible extent. An important method to achieve the objective is by adopting the method of Stratified Random Sampling.

In Stratified random sampling the heterogeneous population which is often encountered in practice is divided into a number of sub-population is called strata. The division of the population into different strata is done in such a way that there is

i) As great homogeneity as possible between the units within each stratum.

ii) There is as marked difference (heterogeneity) as possible between its belonging to different strata.

In brief, the method of stratified sampling can be given as

i) Formation of strata.

ii) Number of strata to be made.

iii) Allocation of sample size within each stratum.

iv) Analysis of data from a stratified design.

We shall discuss the first two points after examining the last two point relating to the theory of stratified sampling.

5.5 ADVANTAGES OF STRATIFIED SAMPLING

The following are the main advantages of stratified sampling:

i) Precision of the estimate is increased i.e. stratified sampling is more efficient than SRS.

ii) Supervision of field work is more convenient and simple.

iii) Selection of the unit is simple. They can be easily located and enumerated.

iv) Non-sampling errors are very much minimized.

v) Stratification ensures adequate representation to various groups of the population, which may be of some interest or importance.

vi) Stratification makes it possible to use different sampling design in different strata.

5.6 DISADVANTAGES OF STRATIFIED SAMPLING

i) Stratification requires to be done carefully and each stratum must contain homogenous items as far as possible. In the absence of this representativeness of the sample cannot be ensured.

ii) Cost per observation may be high.

5.7 MEAN AND VARIANCE OF STRATIFIED SAMPLING

Theorem 1: If sampling is carried out independently and simple random sampling is drawn without replacement from each stratum then $\bar{y}_{st}$ is an unbiased estimation of $\bar{Y}$.

$$V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i}$$

$$= \sum_{i=1}^{k} p_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{S_i^2}{n_i}$$

Where $p_i^2 = \frac{N_i^2}{N^2}$

Proof : we know that

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i \qquad \text{——————} \text{ (i)}$$

Since sampling is done independently in each stratum, we obtain from equation (i)

i.e. $E(\bar{y}_{st}) = E[\frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i]$

$$= \frac{1}{N}\sum_{i=1}^{k} N_i E(\bar{y}_i)$$

$$= \frac{1}{N}\sum_{i=1}^{k} N_i \bar{Y}_i \qquad [\because E(\bar{y}_i) = \bar{Y}_i]$$

$$\Longrightarrow E(\bar{y}_{st}) = \frac{1}{N}\sum_{i=1}^{k} N_i \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij} \qquad [\because \bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}]$$

$$\Longrightarrow E(\bar{y}_{st}) = \frac{1}{N}\sum_{i=1}^{k} \sum_{j=1}^{N_i} Y_{ij}$$

$$\Longrightarrow E(\bar{y}_{st}) = \bar{Y}$$

Hence $\bar{y}_{st}$ is an unbiased estimate of $\bar{Y}$. Now

$$V(\bar{y}_{st}) = E[\bar{y}_{st} - E(\bar{y}_{st})]^2$$

$$= E[\bar{y}_{st} - \bar{Y}]^2$$

$$= E[\frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i - \frac{1}{N}\sum_{i=1}^{k} \sum_{j=1}^{N_i} Y_{ij}]^2$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i - \frac{1}{N}\sum_{i=1}^{k} N_i \bar{Y}_i\right]^2$$

$$\Longrightarrow V(\bar{y}_{st}) = \frac{1}{N^2} E[\sum_{i=1}^{k} N_i (\bar{y}_i - \bar{Y}_i)]^2$$

$$= \frac{1}{N^2} E[\sum_{i=1}^{k} N_i^2 (\bar{y}_i - \bar{Y}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{N_i} N_i N_j (\bar{y}_i - \bar{Y}_i)(\bar{y}_j - \bar{Y}_j)]$$

$$\Longrightarrow V(\bar{y}_{st}) = \frac{1}{N^2}[\sum_{i=1}^{k} N_i^2 E(\bar{y}_i - \bar{Y}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{N_i} N_i N_j E(\bar{y}_i - \bar{Y}_i)(\bar{y}_j - \bar{Y}_j)] \qquad \text{.....(ii)}$$

Since sampling is done independently in each stratum and SRSWOR is drawn from each stratum and we know that

$E(\bar{y}_i) = \bar{Y}_i$; $E(\bar{y}_j) = \bar{Y}_j$

and $V(\bar{y}_i) = E[\bar{y}_i - \bar{Y}_i]^2 = \frac{N_i - n_i}{N_i n_i} S_i^2$

hence we obtain from equation (ii)

$$V(\bar{y}_{st}) = \frac{1}{N^2}\left[\sum_{i=1}^{k} N_i^2 \frac{N_i - n_i}{N_i n_i} S_i^2 + 0\right]$$

[$\because$ y_i and Y_j are independent therefore covariance $= 0$]

$$= \frac{1}{N^2}\left[\sum_{i=1}^{k} N_i^2 \frac{N_i - n_i}{N_i n_i} S_i^2 + 0\right]$$

Multiply with N_i we get

$$V(\bar{y}_{st}) = \frac{1}{N^2}\left[\sum_{i=1}^{k} \frac{N_i^2}{N_i}(N_i - n_i)\frac{S_i^2}{n_i}\right]$$

$$V(\bar{y}_{st}) = \frac{N_i^2}{N^2}\sum_{i=1}^{k} \left(\frac{N_i - n_i}{N_i}\right)\frac{S_i^2}{n_i}$$

$$\Rightarrow P_i^2 \sum_{i=1}^{k} \left(1 - \frac{n_i}{N_i}\right)\frac{S_i^2}{n_i}$$

Hence the theorem is proved.

5.8 SELF ASSESSMENT

Q.1 A random sample of size n is selected from a population containing N units and the sample units are allocated L strata on the basis of information collected above them. Denoting by n_h the number of sample units falling in stratum h, derive the variance of $\sum_{h=1}^{L} \frac{N_h}{N} \bar{Y}_h$.

Q.2 For the sampling scheme in which the population is split at random into substrata containing N_i, i=1, 2, 3 - - - n units and one unit is selected with pp of x from each sub stratum, suggest an unbiased estimator for the population total and derive its variance.

Q.3 Explain the method of stratified sampling. Why is it preferred to SRS?

5.9 SUMMARY

The strata should be non overlapping and should together comprise the whole population. Stratification serves many useful purpose. Stratification makes it possible to use different sampling designs. Stratification is not uniformly available for all units in the population. In these cases, the whole population is subdivided into strata according to the nature of the information available and some suitable sampling scheme of selection of units within these strata is adopted. For stratified random sampling WOR the sample estimate $V(\bar{y}_{st})$ is unbiased and its sampling variance is given by

$$V(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{N_i(N_i-n_i)}{N^2 n_i} S_i^2$$

$$= \sum_{i=1}^{k} \frac{W_i^2(1-f_i)}{n_i} S_i^2$$

Mean of the Stratified sampling is given by

$$\bar{Y} = \sum_{i=1}^{k} \frac{n_i y_i}{n}$$

Variance of the Stratified sampling is given by

$$S_i^2 = \sum_{j=1}^{n_i} \frac{Y_{ij}-\bar{Y}_i^2}{n_i-1}$$

The main objective of the Stratification is to give a better cross-section of the population so as to gain a higher degree of relation precision. To achieve this, the following points are to be examined carefully:

i) Formation of strata to be made.

ii) Number of strata to be made.

iii) Allocation of sample size within each stratum.

iv) Analysis of data from a stratified design.

ALLOCATION PROBLEM

STRUCTURE

6.1 Objectives

6.2 Introduction

6.3 Theorem based on proportional and optimum allocation

6.4 Self Assessment

6.1 OBJECTIVES

- To introduce the proportional allocation.
- To introduce the optimum allocation.
- To introduce the problem based on proportional allocation and optimum allocation.
- To introduce the comparison.

6.2 INTRODUCTION

In stratified sampling, the allocation of the sample to different strata is done by the consideration of three factors viz.

i) The total number of units in the stratum i.e. stratum size.

ii) The variability within the stratum.

iii) The cost in taking observations per sampling unit in the stratum.

A good allocation is one where maximum precision is obtained with minimum resources or, in other words, the criteria for allocation is to minimize the budget for a given variance or minimize the variance for a fixed budget, thus making the most effective use of the available resources.

There are three methods of allocation of sample sizes of the different strata in Stratified sampling procedure. These are:

i) Equal allocation.

ii) Proportional allocation

iii) Optimum allocation.

i) Equal allocation:

If the sample size distributed equally to different strata i.e. $n_1 = n_2 = - - - = n_i$ then allocation is called equal allocation.

ii) Proportional allocation:

Proportional allocation deals with allocating the total sample size n to different strata according to the size of the strata, symbolically,

$n_1 \propto N_1$ where I =1, 2, 3 - - - k

or $n_i = CN_i$ .....(i)

where C is the constant of proportionality summing from i= 1 to k, therefore from equation (i) we get

$$\sum_{i=1}^{k} n_i = C \sum_{i=1}^{k} N_i$$

or n = CN

or $C = \frac{n}{N}$, Hence from equation (i)

$n_i = \frac{n}{N} N_i$ .....(ii)

iii) Optimum allocation:

Optimum allocation also called as Neyman allocation. If the sample is allocated with

n_i proportional to $N_i S_i$ then the allocation is called as Neyman allocation or optimum allocation i.e. $n_i \propto N_i S_i$ where S_i is the standard deviation of the ith stratum.

$$n_i = kN_i S_i \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(i)}$$

Taking summation both sides and taking i= 1, 2, 3 - - - k we get

$$\sum_{i=1}^{k} n_i = k\sum_{i=1}^{k} N_i S_i$$

$$\Rightarrow k = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} N_i S_i}$$

$$\Rightarrow k = \frac{n}{\sum_{i=1}^{k} N_i S_i}$$

Put the value of k in equation (i) we get

$$n_i = \frac{n N_i S_i}{\sum_{i=1}^{k} N_i S_i}, \; i = 1, 2, 3 - - - k$$

6.3 THEOREM BASED ON PROPORTIONAL AND OPTIMUM ALLOCATION

Theorem 1: In a stratified sampling under proportional allocation variance of $\bar{y}_{st}$ is given by

$$V_{prop}(\bar{y}_{st}) = \sum_{i=1}^{k} p_i \frac{S_i^2}{n}(1 - f)$$

$$\Rightarrow V_{prop}(\bar{y}_{st}) = \frac{N - n}{Nn}\sum_{i=1}^{k} p_i {S_i}^2, \text{ where } f = \frac{n}{N} \text{ and } p_i = \frac{N_i}{N}$$

Proof: we know that

$$V(\bar{y}_{st}) = \frac{1}{N^2}\left[\sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i}\right]$$

Moreover under proportional allocation

$$n_i = \frac{n}{N} N_i$$

Thus

$$V_{prop}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{k} N_i\left(N_i - \frac{n}{N}N_i\right)\frac{S_i^2}{\frac{n}{N}N_i}$$

$$= \frac{1}{N^2}\sum_{i=1}^{k}\left(N_i - \frac{n}{N}N_i\right)\frac{S_i^2}{n}N$$

$$= \frac{1}{N}\sum_{i=1}^{k}\left(N_i - \frac{n}{N}N_i\right)\frac{S_i^2}{n}$$

$$= \frac{1}{N}\sum_{i=1}^{k} N_i\left(1 - \frac{n}{N}\right)\frac{S_i^2}{n}$$

$$= \frac{1}{Nn}\sum_{i=1}^{k} N_i(1-f)\,S_i^2$$

$$\Longrightarrow V_{prop}(\bar{y}_{st}) = \frac{1}{n}\sum_{i=1}^{k}\frac{N_i}{N}(1-f)\,S_i^2$$

$$= \frac{1}{n}\sum_{i=1}^{k} p_i\,(1-f)\,S_i^2$$

$$= \sum_{i=1}^{k} p_i\,\frac{S_i^2}{n}(1-f)$$

$$= \frac{(1-f)}{n}\sum_{i=1}^{k} p_i\,S_i^2$$

$$= \frac{\left(1-\frac{n}{N}\right)}{n}\sum_{i=1}^{k} p_i\,S_i^2$$

$$\Longrightarrow V_{prop}(\bar{y}_{st}) = \frac{(N-n)}{Nn}\sum_{i=1}^{k} p_i\,S_i^2$$

Hence Proved

Theorem 2: In simple random sampling without replacement (SRSWOR) under the cost function

$$C = C_0 + \sum_{i=1}^{k} C_i\,n_i$$

The variance $V(\bar{y}_{st})$ is minimum when

$$n_i = \frac{n_i N_i S_i \Big/ \sqrt{C_i}}{\sum_{i=1}^{k} N_i S_i \Big/ \sqrt{C_i}}$$

Proof: we know that

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i}$$

$$= \frac{1}{N^2} \sum_{i=1}^{k} \frac{N_i^2 S_i^2}{n_i} - \frac{1}{N^2}\sum_{i=1}^{k} N_i S_i^2$$

$$= \sum_{i=1}^{k} \frac{N_i^2}{N^2}\frac{S_i^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} \frac{N_i}{N} S_i^2$$

$$\Rightarrow V(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{p_i^2 S_i^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(i)}$$

Now we have to minimize equation (i) subject to the cost function

$$C = C_0 + \sum_{i=1}^{k} C_i n_i$$

Next, we have to use Lagranges method of undermined multiplier to achieve this i.e.

Let

$$\emptyset = V(\bar{y}_{st}) + \lambda \{C_0 + \sum_{i=1}^{k}(C_i n_i - C)\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(ii)}$$

Put the value of $V(\bar{y}_{st})$ from eq. (i) into eq. (ii) we get

$$\emptyset = \sum_{i=1}^{k} \frac{p_i^2 S_i^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2 + \lambda \{C_0 + \sum_{i=1}^{k}(C_i n_i - C)\} \ldots\ldots\ldots\ldots\text{(iii)}$$

Where λ is Lagranges multiplier. Differentiate eq. (iii) with respect to n_i and equating the differentiate coefficient to zero, we get

$$\Rightarrow -\frac{p_i^2 S_i^2}{n_i^2} - 0 + 0 + \lambda C_i = 0$$

$$\Rightarrow -\frac{p_i^2 S_i^2}{n_i^2} + \lambda C_i = 0$$

$$\Longrightarrow \lambda C_i = \frac{p_i^2 S_i^2}{n_i^2}$$

$$\Longrightarrow n_i^2 \lambda = \frac{p_i^2 S_i^2}{C_i} \quad \text{.....(iv)}$$

By taking the square root both sides we get

$$n_i \sqrt{\lambda} = \frac{p_i S_i}{\sqrt{C_i}} \quad \text{.....(v)}$$

Summing equation (v) over i=1, 2, 3, - - - k we obtain

$$\sqrt{\lambda} \sum_{i=1}^{k} n_i = \frac{\sum_{i=1}^{k} p_i S_i}{\sqrt{C_i}}$$

$$\Longrightarrow \sqrt{\lambda}\, n = \frac{\sum_{i=1}^{k} p_i S_i}{\sqrt{C_i}} \quad \text{.....(vi)}$$

Divide equation (v) by equation (vi) we get

$$\frac{n_i}{n} = \frac{\frac{p_i S_i}{\sqrt{C_i}}}{\frac{\sum_{i=1}^{k} p_i S_i}{\sqrt{C_i}}}$$

$$\text{or } n_i = \frac{n\left(\frac{N_i}{N}\right)\frac{S_i}{\sqrt{C_i}}}{\sum_{i=1}^{k}\left(\frac{N_i}{N}\right)\frac{S_i}{\sqrt{C_i}}}$$

$$\text{or } n_i = \frac{nN_i\frac{S_i}{\sqrt{C_i}}}{\sum_{i=1}^{k} N_i\frac{S_i}{\sqrt{C_i}}}$$

Hence proved

Theorem 3: In Stratified sampling, the variance of the estimate) of optimum or Neyman allocation is given by

$$V_{opt}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} p_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2$$

Proof: We know that

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} N_i(N_i - n_i)\frac{S_i^2}{n_i}$$

$$= \frac{1}{N^2} \sum_{i=1}^{k} \frac{N_i^2 S_i^2}{n_i} - \frac{1}{N^2}\sum_{i=1}^{k} N_i S_i^2$$

$$= \sum_{i=1}^{k} \frac{N_i^2}{N^2}\frac{S_i^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} \frac{N_i}{N} S_i^2$$

$$\Longrightarrow V(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{p_i^2 S_i^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\text{(i)}$$

Under optimum allocation

$$n_i = \frac{n N_i S_i}{\sum_{i=1}^{k} N_i S_i} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(ii)}$$

Divide Numerator and Denominator of equation (ii) by N, we get

$$n_i = \frac{n\frac{N_i}{N} S_i}{\sum_{i=1}^{k} \frac{N_i}{N} S_i} \quad \Rightarrow n_i = \frac{n p_i S_i}{\sum_{i=1}^{k} p_i S_i} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(iii)}$$

Making substitution from equation (iii) to equation (i) we get

$$V_{opt}(\bar{y}_{st}) = \frac{\sum_{i=1}^{k} p_i^2 S_i^2}{n p_i \frac{S_i}{\sum_{i=1}^{k} p_i S_i}} - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2$$

$$= \frac{\sum_{i=1}^{k} p_i^2 S_i^2 \sum_{i=1}^{k} p_i S_i}{n p_i S_i} - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2$$

$$V_{opt}(\bar{y}_{st}) = \frac{1}{n}(\sum_{i=1}^{k} p_i S_i)^2 - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2$$

Theorem 4: Prove that

$$V_{opt}(\bar{y}_{st}) \leq V_{prop}(\bar{y}_{st}) \leq V_{ran}(\bar{y})$$

Proof:

$$V_{ran}(\bar{y}) = \frac{N-n}{N}\frac{S^2}{n}$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)S^2 \approx \frac{S^2}{n}$$

$$\left[\because \frac{1}{N} = 0, \text{when N is large}\right]$$

also

$$V_{prop}(\bar{y}_{st}) = \frac{1}{n}\sum_{i=1}^{k} p_i S_i^2 - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2$$

$$V_{prop}(\bar{y}_{st}) = \frac{1}{n}\sum_{i=1}^{k} p_i S_i^2 \quad \text{.....(ii)}$$

and

$$V_{opt}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} p_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2$$

$$V_{opt}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k} p_i S_i\right)^2 \quad \text{.....(iii)}$$

Also we know that

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{k} \sum_{j=1}^{N_i}(Y_{ij} - \bar{Y})^2$$

$$\Rightarrow (N-1)S^2 = \sum_{i=1}^{k} \sum_{j=1}^{N_i}(Y_{ij} - \bar{Y})^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i + \bar{Y}_i - \bar{Y})^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{N_i}\{(Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \bar{Y})\}^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{N_i}(\bar{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{k} \sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \bar{Y})$$

$$\Rightarrow (N-1)S^2 = \sum_{i=1}^{k} \sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{k} N_i(\bar{Y}_i - \bar{Y}) + 0$$

[Here "j" is constant]

[Note: The product vanishes since $\sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i)$ being the algebraic sum of deviations from mean is zero. In brief, the cross product term is zero because the sum of the deviation taken from A.M. is always equal to zero]

$$\Rightarrow (N-1)S^2 = \sum_{i=1}^{k}(N_i - 1)S_i^2 + \sum_{i=1}^{k} N_i(\bar{Y}_i - \bar{Y})^2 \text{.....(iv)}$$

[because $S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i)^2$]

Taking N_i to be large so that $N_i - 1 \sim N$, we obtain from equation (iv)

$$NS^2 = \sum_{i=1}^{k} N_i S_i^2 + \sum_{i=1}^{k} N_i(\bar{Y}_i - \bar{Y})^2$$

$$\Rightarrow S^2 = \frac{1}{N}\sum_{i=1}^{k} N_i S_i^2 + \frac{1}{N}\sum_{i=1}^{k} N_i(\bar{Y}_i - \bar{Y})^2$$

$$\Rightarrow S^2 = \sum_{i=1}^{k} \frac{N_i}{N} S_i^2 + \sum_{i=1}^{k} \frac{N_i}{N}(\bar{Y}_i - \bar{Y})^2$$

Divide both sides by 'n' we get

$$\frac{S^2}{n} = \frac{1}{n}\sum_{i=1}^{k} p_i S_i^2 + \frac{1}{n}\sum_{i=1}^{k} p_i(\bar{Y}_i - \bar{Y})^2$$

using equation (i) and (ii) we get

$$V_{ran}(\bar{y}) = V_{prop}(\bar{y}_{st}) + \frac{1}{n}\sum_{i=1}^{k} p_i(\bar{Y}_i - \bar{Y})^2$$

$$\Rightarrow V_{ran}(\bar{y}) = V_{prop}(\bar{y}_{st}) + \text{a positive quantity}$$

$$\Rightarrow V_{prop}(\bar{y}_{st}) \leq V_{ran}(\bar{y}) \text{.....(v)}$$

Subtracting equation (iii) from (ii) we get

$$V_{prop}(\bar{y}_{st}) - V_{opt}(\bar{y}_{st}) = \frac{1}{n}[\sum_{i=1}^{k} p_i S_i^2 - (\sum_{i=1}^{k} p_i S_i)^2 \text{.....(vi)}$$

Let us suppose that

$\sum_{i=1}^{k} p_i S_i = S_w$

$\Rightarrow \sum_{i=1}^{k} p_i (S_i - S_w)^2 = \sum_{i=1}^{k} p_i (S_i^2 + S_w^2 - 2S_i S_w)$

$\Rightarrow \sum_{i=1}^{k} p_i (S_i - S_w)^2 = \sum_{i=1}^{k} p_i S_i^2 - 2S_w^2 \sum_{i=1}^{k} p_i S_i^2 + S_w^2 \sum_{i=1}^{k} p_i$

$\Rightarrow \sum_{i=1}^{k} p_i (S_i - S_w)^2 = \sum_{i=1}^{k} p_i S_i^2 - 2S_w^2 + S_w^2$

Note: because $\sum_{i=1}^{k} p_i S_i = S_w$ and $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} \frac{N_i}{N} = \frac{1}{N}\sum_{i=1}^{k} N_i = \frac{N}{N} = 1$

$\Rightarrow \sum_{i=1}^{k} p_i (S_i - S_w)^2 = \sum_{i=1}^{k} p_i S_i^2 - S_w^2$

$= \sum_{i=1}^{k} p_i S_i^2 - (\sum_{i=1}^{k} p_i S_i)^2$ .....(vii)

[because $\sum_{i=1}^{k} p_i S_i = S_w$]

Using equation (vii) we obtain from equation (vi)

$$V_{prop}(\bar{y}_{st}) - V_{opt}(\bar{y}_{st}) = \frac{1}{n}[\sum_{i=1}^{k} p_i (S_i - S_w)^2] \geq 0$$

As we know that square quantity is always positive

$\Rightarrow V_{prop}(\bar{y}_{st}) - V_{opt}(\bar{y}_{st}) \geq 0$

$\Rightarrow V_{prop}(\bar{y}_{st}) \geq V_{opt}(\bar{y}_{st})$.....(viii)

combine equation (v) and (viii) we get

$V_{ran}(\bar{y}_{st}) \geq V_{prop}(\bar{y}_{st}) \geq V_{opt}(\bar{y}_{st})$

$\Rightarrow V_{opt}(\bar{y}_{st}) \leq V_{prop}(\bar{y}_{st}) \leq V_{ran}(\bar{y})$

Hence proved

6.4 SELFASSESSMENT

Q.1 Define in brief equal allocation.

Q.2 Define proportional allocation.

Q.3 Define optimum allocation.

Q.4 Derive the variance of the estimator under proportional allocation.

6.4 SUMMARY

In simple random sampling, it has been seen that the precision of the standard estimator of the population total (mean) depends on two aspects namely the sample size and the variability of the character under study. Therefore, in order to get an estimator with increased precision one can increase the sample size. However, consideration of cost limit the size of the sample. The other possible way to estimate the population total with greater precision is to divide the population into several groups. Dependent on the nature of the strata, different sampling design can be used in different strata. for example, in the absence of suitable size information SRS can be used in few strata, whereas probability proportional to size sampling can be used in the remaining strata when size information is available in those strata.

Also in this lesson, we made a comparative study of the usual estimators under simple random sampling without stratification and stratified random sampling employing various schemes of allocation i.e. proportional and optimum allocation the variances of these estimator of mean are denoted by $V_{opt}(\bar{y}_{st}), V_{prop}(\bar{y}_{st})$ and $V_{ran}(\bar{y})$ respectively.

SYSTEMATIC SAMPLING

STRUCTURE

7.1 Objectives

7.2 Introduction

7.3 Systematic sampling

7.4 Advantages and disadvantages of Systematic sampling

7.5 Mean and variance of Systematic sampling

7.6 Self Assessment

7.7 Summary

7.1 OBJECTIVES

- To introduce the Systematic sampling.
- To introduce the mean of Systematic sampling.
- To introduce the variance of Systematic sampling.
- To introduce the problems and self assessment based upon the Systematic sampling.

7.2 INTRODUCTION

In this lesson, collection of sampling schemes called systematic sampling schemes which have several practical advantages are considered. In these schemes instead of selecting n units at random, the sample units are decided by a single number chosen at random.

Consider s finite population of size N, the units of which are identified by the labels 1, 2, 3 - - - N and ordered in ascending order according to their labels. Unless otherwise mentioned, it is assumed that the population size N is expressible as product of the sample size 'n' and some positive integer k, which is known as the reciprocal of the sampling interval.

7.3 SYSTEMATIC SAMPLING

Often we have to obtain the information from the cards or registers which are in serial order. Sometimes we might need the sample of trees from a forest or houses in a city. In such a case, a sample plan known as systematic sampling often works better than the SRS. In brief, a collection of sample schemes are called systematic sample which have several practical advantages are considered. For example, if there are 100 units in a population serially numbered from 1 to 100 and we want to draw a sample of 5 units. Draw a random number from 1 to 20 since k = 20 and let the selected number is 7 i.e. j = 7. Then select units in serial number 7, 27, 47, 67, 87. These units constitute a systematic sample of 5 units.

There are two methods in systematic sampling

i. Linear systematic sampling

ii. Circular systemic sampling.

i. Linear Systematic Sampling

Suppose a population consist of N units and from this a systematic sample of n units is to be selected. Also assume that N is a multiple of n i.e. N = nk. The procedure is to select a random number, say j, such that $1 \leq j \leq k$ and select the jth and every subsequent j + k, j + 2k - - - j + (n + 1) kth positional units. This sampling plan is known as linear systematic sampling. For example if N = 13, n = 4, then k has to be taken as 4, now select a random sample from 1 to 4, say it is 3, then the remaining three units to be selected are at positions 7, 11, 15. There is no unit in the population at serial number 15. Hence we can select only a sample of size 3 in this situation. To cope with this situation, a sampling plan known as circular systematic sampling is to be used.

(ii) Circular systematic Sampling :

Circular systematic is applied when $N \neq nk$. This type of sampling was first used by D. B. Lahri (1952) in national sampling surveys. Here, we take N/n as k by rounding of N/n to the nearest integer . Select a random number from 1 to N. Let this number be m. Now select every $(m + jk)^{th}$ unit when $m + jk < N$ and select every $(m + jk - N)^{th}$ unit when $m + jk > N$ putting j= 1, 2, 3, - - - till n units are selected. By this method we always get a sample of size n. Suppose for the example with N = 13, n = 4 and k = 3 the randomly selected number from 1 to 13 is 8 and later the 11^{th} , 1^{st} , 4^{th} and 7^{th} units are selected. When N=nk, the linear and circular systematic plans becomes identical. For example, if N = 12, n = 4, k = 3 and the selected units from 1 to 3 is 3 and from 1 to 12 is either of the number 3, 6, 9 and 12. Also if the selected number from 1 to 3 is 2 and from 1 to 12 is any one of the 2, 5, 8 or 11 the selected units will be those bearing the serial numbers 2, 5, 8, and 11. This example shows the equivalence between two selection plans.

7.4 ADVANTAGES AND DISADVANTAGES OF SYSTEMATIC SAMPLING

Systematic sampling has some advantages as well as disadvantages some of the advantages are:

i. The method of selection is very simple and not very expensive.

ii. The sample is evenly distributed over the whole population and hence all contiguous parts of the population are represented in the sample.

iii. It has an advantage over sampling plan because it has organizing control of field work.

The disadvantages are:

i. If the variation in the units is periodic i.e. the units at regular intervals are correlated then the sample becomes highly biased. For example, if the houses are in blocks and a corner house is selected at random then all other houses selected in the sample will be corner ones and this definitely give a biased sample.

ii. No single reliable formula is good enough, if the population is of the type it has been assumed to be. This is of course, a great draw back.

7.5 MEAN AND VARIANCE OF SYSTEMATIC SAMPLING

Let the observations on selected units be $x_1, x_2, x_3 - - - x_n$. Then the sample means $\bar{x}_{sy}$ of the systematic sample is

$$\bar{x}_{sy} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

for i = 1, 2, 3 - - - n and variance of when N = nk is

$$V(\bar{x}_{sy}) = \frac{N-1}{N} S^2 - \frac{K(n-1)}{N} S^2_{wsy} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

As we have noticed that for every random number selected from 1 to k, a separate systematic sample will be obtained. Let represent the i^{th} observation in the j^{th} systematic sample which is obtained when a random number j= 1, 2, 3 - - - k is selected. Thus in equation (2), the variance between systematic sample is

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - \mu)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

For j = 1, 2, 3 - - - k and I = 1, 2, - - - n

$$\text{also } \frac{N-1}{N} S^2 = \sigma^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

and the variance S^2_{wsy} within the systematic sample is

$$S^2_{wsy} = \frac{1}{k(N-1)}\sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - \bar{x}_j)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

and

$$\frac{k(N-1)}{N} S^2_{wsy} = \frac{1}{N}\sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - \bar{x}_j)^2 = \sigma^2_w \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

Thus, the formula in equation (2) reduces to

$$V(\bar{x}_{sy}) = (\sigma^2 - \sigma^2_w) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(7)$$

σ^2, the population variance is a constant quantity and is the variance within the systematic sample which is being subtracted from to get . Hence, greater is the heterogeneity among units within sample, the better it is.

Remarks:

If the units at equal intervals are more alike, then the systematic sample is not good sampling scheme.

7.6 SELFASSESSMENT

Q.1. Derive the variance of the conventional estimator in SRS. Linear systematic sampling and stratified sampling assuming $Y_i = a + b_i + c_i^2$ and compare them.

Q.2. Derive the variance of conventional estimator for the population mean under circular systematic sampling assuming the presence of linear trend.

Q.3. Develop correction estimator for the population mean in the presence of parabolic trend assuming the sample size is odd under linear systematic sampling.

7.4 SUMMARY

In linear systematic sampling the second order inclusion probabilities are not positive for all pairs of units in the population. Assuming that the two units have been selected with simple random sampling without replacement from the 2k units in the i^{th} term in the brackets on the right hand side will be given by

$$\frac{k-1}{k}\left\{\frac{Y_{2i} - Y_{2i-1}}{2}\right\}^2$$

They have proved that if U + vd is less than or equal to N then the above sampling scheme will yield distinct units and the second order inclusion probabilities are positive. They have observed that in situations where usual systematic sampling performs better than simple random sampling and suggested procedure also leads to similar results and for some situations it provides better results than linear systematic sampling.

CLUSTER SAMPLING

STRUCTURE

8.1 Objectives

8.2 Introduction

8.3 Equal Cluster sampling

8.4 Cluster sampling with unequal cluster size

8.5 Varying probability Cluster sampling

8.6 Efficiency of Cluster sampling

8.7 Self Assessment

8.8 Summary

8.1 OBJECTIVES

- To introduce the definition and introduction of Cluster sampling.
- To introduce the Cluster sampling in equal sampling in equal cluster size.
- To introduce Cluster sampling in unequal cluster size.
- To introduce theorems and varying probability Cluster sampling.
- To introduce the problems and self assessments based upon the cluster sampling.

8.2 INTRODUCTION

In random sampling, it is presumed that the population has been divided into a finite number of distinct and identifiable units defined as sampling units. The smallest unit into which the population can be divided is called an element of population. A group of such elements is known as cluster. When the sampling unit is a cluster, then the procedure is called cluster sampling. If the entire area containing the population under study is divided into smaller segments and each element in the population belongs to one and only one segment, the procedure in the population sometimes called area sampling.

Generally, identification and location of an element requires considerable time. However, once an element has been located, the time taken for surveying a few neighboring element is small. Thus the main function of Cluster sampling is to specify clusters or to divide the population into approximate clusters. Clusters are generally made up of neighboring elements and therefore the elements within a cluster tend to have similar characteristic. As a simple rule, the number of elements in a cluster should be small and the number of clusters should be large. After dividing the population into specified clusters. The required numbers of clusters can be selected either by equal or unequal probabilities of selection. All the elements in selected clusters are enumerated.

For a given number of sampling units, cluster sampling is more convenient and less costly. The advantages of cluster sampling are that:

i) Collection of data for neighboring elements is easier, cheaper, faster and operationally more convenient than observing units spread over a region.

ii) It is less costly than simple random sampling due to the saving of time in journeys, identification, contacts etc.

iii) When the sampling frame of elements may not be readily available.

Even if such a frame is made available it would be expensive too base an enquiry on a simple random sample of elements. From the point of view of Statistical efficiency of units in a Cluster to be similar. In fact, the efficiency of cluster sampling is likely to decrease with increase in cluster size. For a given sample size a smaller sampling unit bring more precise results than a larger sampling unit. In most of the practical situations, the loss in efficiency

may be balanced by the reduction in cost. Therefore, the efficiency per unit cost may be more in cluster sampling than in SRS.

The selection of clusters can be random or by first selecting a unit called a key unit at random and then randomly taking the required number of neighboring units to form the clusters. For example, to estimate the milk production a cluster of three villages can be formed by first selecting a key village at random and then taking two more villages from a block of some specified area. Cluster may be overlapping or non overlapping. Various sampling procedures viz. simple random, Stratified or systematic sampling by treating the clusters themselves a sampling units.

8.3 EQUAL CLUSTER SAMPLING

We shall first consider the case of equal cluster. Suppose the population consists of V clusters each of M elements and that a sample of n clusters is drawn by the method of SRS.

N = Number of clusters in the population

n = Number of clusters in the sample

M = Number of elements in the cluster

Y_{ij} = The value of the characteristics under study for the j^{th} elements(j=1, 2, 3 - - - M) in the i^{th} cluster (i = 1, 2, 3, - - - N)

$\bar{Y}_i$ = $\sum_{j=1}^{M} \frac{Y_{ij}}{M}$ = the mean per element of i^{th}cluster

$\bar{Y}_n$ = $\sum_{i=1}^{N} \sum_{j=1}^{M} \frac{Y_{ij}}{NM}$ = the mean of per element in the population

S_i^2 = $\frac{1}{M-1}\sum_{j=1}^{M}(Y_j - \bar{Y}_i)^2$ = the mean square between elements within the i^{th} cluster (i = 1, 2, 3, - - - N)

S_w^2 = $\frac{1}{N-1}\sum_{j=1}^{N}(\bar{Y}_i - \bar{Y}_N)^2$ =the mean square between elements in the population

$\bar{\rho}$ = $\frac{E(Y_{ij}-\bar{Y})}{E(Y_{ij}-\bar{X})} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij}-\bar{Y})(Y_{jk}-\bar{Y})}{(M-1)(NM-1)S^2}$ =intra class correlation coefficient between elements within clusters

Cluster sampling is used when:

i) The sampling frame is not available and it is too expensive and time consuming to prepare it.

ii) The sampling units are situated apart. In this connection selection of elementary units make the survey very cumbersome. For instance, selection of frame in a State.

iii) The elementary units may not be easily identifiable e.g. the animals of certain species, the migratory population etc.

8.4 CLUSTER SAMPLING WITH UNEQUAL CLUSTER SIZE

$\bar{M} = \frac{1}{N}\sum_{i=1}^{N} M_i$: average size of the clusters

$\bar{Y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} Y_{ij}$: number of clusters in the sample

$\bar{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{N} \bar{Y}_i$: mean of the cluster

$U_{ij} = \frac{M_i Y_{ij}}{\bar{M}}$

$\bar{U}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} U_{ij} = \frac{1}{M_i}\sum_{j=1}^{M_i} Y_{ij} = \frac{M_i Y_{ij}}{\bar{M}} = \frac{1}{\bar{M}}\sum_{j=1}^{M_i} Y_{ij} = \frac{M_i \bar{Y}_i}{\bar{M}}$

$V_i = \frac{M_i}{\bar{M}}$

$\bar{Y} = \frac{1}{\bar{M}N}\sum_{j=1}^{N} M_i \bar{Y}_i = \frac{1}{N}\sum_{i=1}^{N} \bar{U}_i = \bar{\bar{U}}$

Also

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} \bar{V}_i = \frac{1}{N}\sum_{i=1}^{N} \frac{M_i}{\bar{M}} = \frac{N\bar{M}}{N\bar{M}} = 1$

Several sampling strategies are available to estimate the population mean $\bar{\bar{Y}}$ where cluster are of unequal sizes.

a) SRSWOR: The most frequently used estimators are as follows

i) Mean Estimator:

$$\hat{\bar{\bar{Y}}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$$

Where y_{ij} is the jth unit in the i cluster selected in a sample of size n,

where i = 1, 2, 3 - - - n and j = 1, 2, 3 - - - M_i

and

$$\bar{y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij}$$

The estimator $\hat{\bar{Y}}$ is a biased estimator of , the bias being given by

$$\text{Bias}(\hat{\bar{Y}}) = E(\hat{\bar{Y}}) - \bar{Y}$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(\bar{y}_i) - \bar{Y}$$

$$\Longrightarrow \text{Bias}\left(\hat{\bar{Y}}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i - \bar{Y}$$

$$\Longrightarrow \text{Bias}\left(\hat{\bar{Y}}\right) = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i - \frac{1}{N\bar{M}}\sum_{i=1}^{N} M_i - \bar{Y}_i$$

$$= \frac{1}{N\bar{M}}\sum_{i=1}^{N}(M_i\bar{Y}_i - \bar{M}\bar{Y}_i)$$

$$= \frac{1}{N\bar{M}}\sum_{i=1}^{N}\bar{Y}_i(M_i - \bar{M})$$

$$V(\bar{Y}) = (\frac{1}{n} - \frac{1}{N})S_b^2$$

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

An unbiased estimate of $V\left(\hat{\bar{Y}}\right)$ is

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

Where

$$S_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{\bar{y}})^2$$

i) **Unbiased Estimator**

An unbiased estimator of $\bar{Y}$ is $\hat{\bar{Y}} = \bar{\bar{U}}$,

where $\bar{\bar{U}} = \frac{M_i Y_{ij}}{\bar{M}}$

the expected value of $\overline{\overline{U}}$ is

$$E(\overline{\overline{U}}) = \frac{1}{n}\sum_{i=1}^{n} E(\overline{U}_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{N}\sum_{i=1}^{N} \overline{U}_i$$

$$= \frac{1}{N}\sum_{i=1}^{N} \overline{U}_i = \overline{Y}$$

For the unbiased estimator $\overline{\overline{U}}$

$$V(\overline{\overline{U}}) = (\frac{1}{n} - \frac{1}{N})S_U^2$$

where

$$S_U^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\frac{M_i\overline{Y}_i}{M} - \overline{Y})^2$$

An unbiased estimator of $V(\overline{\overline{U}})$ is given by

$$\hat{V}(\overline{\overline{U}}) = (\frac{1}{n} - \frac{1}{N})S_U^2$$

where

$$S_U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\overline{U}_i - \overline{\overline{U}})^2$$

8.5 VARYING PROBABILITY CLUSTER SAMPLING

In many practical situations, cluster size is positively correlated with the variable under study. In these cases, it is advisable to select the clusters with probability proportional to the number of elements in the cluster.

Theorem: If a sample of n-clusters is drawn with probabilities P_i and with replacement, then an unbiased estimator of $\overline{Y}$ is given by

$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} \overline{Z}_i$ with its sampling variance

$$V(\overline{Z}_n) = \frac{1}{n}\sum_{i=1}^{n} P_i(\overline{Z}_i - \overline{Y})^2$$

Proof:- To Prove that $\bar{Z}_n$ is unbiased, we have

$$E(\bar{Z}_n) = E[\frac{1}{n}\sum_{i=1}^{n} \bar{Z}_i] = \frac{1}{n}\sum_{i=1}^{n} E(\bar{Z}_n)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \sum_{i=1}^{N} P_i \frac{M_i \bar{Y}_i}{M_{ij} P_i}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \sum_{i=1}^{N} P_i \frac{M_i \bar{Y}_i}{M_o}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \bar{Y} = \bar{Y}$$

To obtain the sampling variance of $\bar{Z}_n$ we have

$$V(\bar{Z}_n) = \frac{1}{n}\sum_{i=1}^{n} P_i (\bar{Z}_i - \bar{Y})^2$$

Corollary 1: If units are drawn with probabilities proportional to size $P_i = \frac{M_i}{M_o}$ with replacement, then as unbiased estimator of $\bar{Y}$ is given by

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} \bar{Y}_i$$

With variance

$$V(\bar{Z}_n) = \sum_{i=1}^{n} \frac{M_i}{M_o} (\bar{Y}_i - \bar{Y})^2$$

Corollary 2: If units are drawn with probability P_i with replacement an unbiased estimator of variance $V(\bar{Z}_n)$ is given by

$$V(\bar{Z}_n) = \frac{\sum_{i=1}^{n} (\bar{Z}_i - \bar{Z}_n)^2}{n(n-1)}$$

8.6 EFFICIENCY OF CLUSTER SAMPLING

We observe that the estimator $\bar{\bar{y}}$ is based on a sample of nM elements. Hence the relative

efficiency of w.r.t. based on a sample of nM elements drawn by SRSWOR from NM elements in the population is relative efficiency i.e.

$$\text{R.E.}(\bar{\bar{y}}) = \frac{\frac{NM-nM}{NMnM}S^2}{\frac{N-n}{Nn}S_b^2} = \frac{\frac{M(N-n)}{MM}}{(N-n)S_b^2}$$

$$\text{R.E.}(\bar{\bar{y}}) = \frac{S^2}{MS_b^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

It follows from the equation (1) that the efficiency of cluster sampling increases as the mean square between clusters decreases. Further

$$(NM-1)S^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij}-\bar{Y})^2$$

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

$$= \sum_{i=1}^{N}\sum_{j=1}^{M}[(Y_{ij}-\bar{Y}_i)^2+2(Y_{ij}-\bar{Y}_i)(\bar{Y}_i-\bar{Y})+(\bar{Y}_i-\bar{Y})^2]$$

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

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

$$= \sum_{i=1}^{N}(M-1)S_i^2+0+M\sum_{i=1}^{N}(\bar{Y}_i-\bar{Y})^2$$

$$\Rightarrow (NM-1)S^2 = (M-1)NS_w^2+M(N-1)S_b^2$$

$$\Rightarrow MS_b^2 = \frac{1}{N-1}[(NM-1)S^2-N(M-1)S_w^2] \quad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

It follows that S_b^2 decreases as S_w^2 increases and hence the efficiency of cluster sampling increases as the mean square within the cluster increases. This suggest that for cluster sampling to be efficient, the cluster should so formed that the variation between the cluster means is as small as possible while the variation with in cluster is as large as possible.

To see how the efficiency of the cluster $\bar{\bar{y}}$ in cluster sampling changes with the size of the cluster. It is convenient to express the variance in terms of intra-class correlation coefficients between elements of the cluster. Now, the intra-class correlation coefficient is defined as

$\rho_c = \frac{cov(y_{ij}, y'_{ij})}{var(y_{ij})}$; where $j = j'$ .....(3)

$$cov(y_{ij}, y'_{ij}) = E[cov\left(\frac{y_{ij}, y'_{ij}}{i}\right) + cov[E\left(\frac{y_{ij}}{i}\right)][E\left(\frac{y_{ij'}}{i}\right)] \text{.....(4)}$$

Where i = 1, 2, 3 - - - N

$$\Rightarrow cov(y_{ij}, y'_{ij}) = \frac{1}{N(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \bar{Y}_i) + (Y'_{ij} - \bar{Y}_i) \text{ ..... (5)}$$

$$\therefore \sum_{i=1}^{N} (Y_{ij} - \bar{Y}_i)^2 = 0$$

$$\text{and } \sum_{j=1}^{M} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \bar{Y}_i)(Y'_{ij} - \bar{Y}_i) = 0$$

$$\text{or } \sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \bar{Y}_i)(Y'_{ij} - \bar{Y}_i) = 0 - \sum_{j=1}^{M} (Y_{ij} - \bar{Y}_i)^2$$

$$\Rightarrow \sum_{i=1}^{N} \sum_{j=1}^{M} (Y_{ij} - \bar{Y}_i)(Y'_{ij} - \bar{Y}_i) = -(M-1)S_i^2 \text{.....(6)}$$

$$\therefore Cov[E(\frac{y_{ij}}{i}) E(\frac{y'_{ij}}{i}) = V(\bar{Y}_i) \text{ .....(7)}$$

We obtain from equation 4, 5, 6 and 7

$$cov(Y_{ij}, Y'_{ij}) = E\left[\frac{\frac{(M-1)}{M(M-1)} S_i^2}{i}\right] + V(\bar{y}_i)$$

$$= \frac{1}{M} E\left[\frac{S_i^2}{i}\right] + V(\bar{y}_i)$$

$$= -\frac{1}{M}\frac{1}{N} \sum_{i=1}^{N} S_i^2 + E(Y_i - \bar{Y}_i)^2$$

$$= -\frac{1}{M} S_w^2 + \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y}_i)^2$$

$$\Rightarrow cov(y_{ij}, y'_{ij}) = -\frac{1}{M} S_w^2 + \frac{N-1}{M} S_b^2 \text{.....(8)}$$

Where $V(Y_{ij}) = E(Y_{ij} - \bar{Y})^2$

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

$$\Longrightarrow V(Y_{ij}) = \frac{NM-1}{NM}S^2 \text{.....(9)}$$

Making substitutions for (8) and (9) to equation (3) we get

$$\boldsymbol{\delta_c} = \frac{\frac{(N-1)}{N}S_b^2 - \frac{S_w^2}{M}}{\frac{NM-1}{NM}S^2} \text{.....(10)}$$

Also from equation (2) we get

$$(N-1)S_b^2 = \frac{1}{M}[(NM-1)S^2 - N(M-1)S_w^2] \text{.....(11)}$$

From equation (10) and (11) we get

$$\boldsymbol{\delta_c} = \frac{\frac{1}{NM}[(NM-1)S^2 - N(M-1)S_w^2] - \frac{S_w^2}{M}}{\frac{NM-1}{NM}S^2}$$

or $(NM-1)S^2 - N(M-1)S^2 - NS_w^2 = (NM-1)S^2\boldsymbol{\delta_c}$

$$\Longrightarrow S_w^2 = \frac{1}{NM}[(NM-1)S^2(1-\boldsymbol{\delta_c}) \text{ .....(12)}$$

From equation (10) and (12) we get

$$\boldsymbol{\delta_c} = \frac{\frac{(N-1)}{N}S_b^2 - \frac{1}{NM^2}[(NM-1)S^2(1-\boldsymbol{\delta_c}) }{\frac{NM-1}{NM}S^2}$$

$$\Longrightarrow S_b^2 = \frac{(NM-1)S^2[1+(M-1)\boldsymbol{\delta_c}]}{M^2(N-1)} \text{ .....(13)}$$

Using the expression (13) for S_b^2 we obtain from equation (1)

$$\text{R.E. } (\bar{\bar{y}}\ ) = \frac{M(N-1)}{(NM-1)}\ \ \frac{1}{[1+(M-1)\boldsymbol{\delta_c}]} \text{.....(14)}$$

We observe from equation (14) that smaller the value of δ_c larger is the efficiency of the cluster sampling as compared to the SRSWOR. Further more

R.E. $(\bar{\bar{y}}\ ) = 1$ when

$$\frac{M(N-1)}{(NM-1)} \quad \frac{1}{[1+(M-1)\delta_c]} = 1$$

$$\Rightarrow (NM-1) + (NM-1)(M-1)\delta_c = M(N-1)$$

$$\Rightarrow \delta_c = \frac{1}{(MN-1)}$$

hence if

$$\Rightarrow \text{if } \delta_c < \frac{1}{(MN-1)} \text{ then } V(\bar{\bar{y}}) = V(\bar{y})$$

$$\text{and if } \delta_c > -\frac{1}{(MN-1)} \text{ then } V(\bar{\bar{y}}) = V(\bar{y})$$

8.7 SELF ASSESSMENTS

(i) If NM elements in a population are grouped at random to form N clusters of M elements each, show that a random sample WOR of n clusters would have the same efficiency as sampling nM elements in random sample WOR.

(ii) A Survey on pepper was conducted to estimate the numbers of people standards and production of pepper in Kerala state (India). For this 3 clusters from 95 were selected by SRSWOR. The formation of the number of pepper standards recorded as given below:

Cluster No.	Cluster size	No. of pepper standard
1	12	41, 16, 19, 15, 144, 212, 57, 2876, 199
2	12	39, 70, 38, 37, 161, 38, 27, 219, 46, 128, 30, 26
3	7	115, 59, 120

Estimate the total number of pepper standards along with standard error for the region, given $\bar{M}$ the average cluster size for the population to be 10.

8.1 SUMMARY

The selection of the cluster can be random or by first selecting a unit called a key unit at random and then randomly taking the required number of neighboring units to form the cluster. For example, to estimate the milk production a cluster of three villages can be formed by first selecting a key village at random and then taking two more villages from a block of some specified area. Cluster may be over lapping or non-overlapping. Various sampling procedures viz. simple random sampling, stratified or systematic sampling procedures can be applied to cluster sampling by treating the cluster themselves as sampling units.

ANALYSIS OF VARIANCE

STRUCTURE

9.1 Objectives

9.2 Introduction

9.3 Sum of squares and Degree of freedom.

9.4 Variability with in classes and Between classes

9.5 Application of Analysis of variance

9.6 Basic assumptions

9.7 Limitations and Precautions of analysis of variance

9.8 Technique in one-way Analysis of variance

9.9 Technique in two-way Analysis of variance

9.10 Self Assessments

9.11 Summary

9.1. OBJECTIVES

The main objectives of this lesson are :

- To introduce the need of analysis of variance.
- To introduce the meaning of analysis of variance

- To provide the applications of analysis of variance
- To provide the fundamental assumptions of analysis of variance
- To introduce the concept of classification.

9.2. INTRODUCTION

Analysis of variance has been defined as the statistical technique for the "separation of variation due to a group of causes from the variation due to a group of causes from the variation due to other groups." Here, when we note the observations from an experiment pertaining to yield or measurement of any other character, we find that the observations vary from one another greatly. This variation is due to a number of factors known as sources of variation and the portions of variation caused by different sources are known as components of variation. The statistical analysis aims at assessing this total variation factor responsible for the same. The analysis of variance is a simple arithmetical process of sorting out the components of variation in a given data.

Analysis of variance is a technique for testing the simultaneous significance of the difference among the mean of several categories (often called classes or groups) as developed by Prof. R.A. fisher in 1920's. According to him "Separation of variance ascribable to one group of causes from the variance ascribable to other groups was the form of this concept." Formally analysis of variance may be defined as :

Analysis of variance is a technique of separating the total variation observed in a set of observed data into its non-negative components, each of which measures the variability ascribable to different specific sources (or factors) and then estimating the variance of the population by independent estimates followed by a comparison of variation due to each assignable factor with that of the chance factor."

A cause of variation which is assignable is called assignable cause or factor of variation which can be specified.

The chance factor is a combination of a large number of small uncontrolled independent causes which cannot be traced out separately.

The analysis of variance technique is based on two principles:

(i) The partitioning of the sum of squares.

(ii) The estimation of the variance of the population by two or more independent estimates followed by a comparison of these estimates where F-test is applied to test the homogeneity of the observations through these estimates.

9.3 SUM OF SQUARES AND DEGREE OF FREEDOM

Before we start learning the principles of analysis of variance, it is essential to know the relation between the variance, sum of squares and the degree of freedom.

(i) Sum of squares :

It means the sum of squares of deviations of the variates from their mean. If X_1, $X_{2....}$,Xn are n variates and $\bar{X}$ is their mean, then sum of **squares of deviations of the variables from their mean** $= \sum_{i=1}^{N} (X - \bar{X})^2$

If zero takes as the arbitrary mean, the deviations of the variates from zero will be the variates themselves, therefore using the short cut method :

Sum of squares $= (X_1^2 + X_2^2 + X_n^2) - (\sum X)^2/n$

(ii) Degrees of freedom :

The number of degrees of freedom is one less than the number of variates in the simple concern. Here it will be (n-1). If the number of treatments is t , the degree freedom for treatment means will be (t-1). In general, the number of degrees of freedom for any number of group means is one less than the number of groups concerned.

(iii) Variance :

It is obtained by dividing the sum of squares by the degrees of freedom, i.e.

Variance = S.S/D.F

9.4 VARIABILITY WITH IN CLASSES AND BETWEEN CLASSES

The variation among the observations of each specific class is called its internal variation and the totality of the interval variations is called variability with in classes.

The totality of variation within each class reflects chance variation under the assumption that the variation due to classes is equal to the total variation and is often called the experimental error. This variation is due to the uncontrolled and non-specific factors.

The totality of variations from one class to another, i.e., variation due to classes is called variability between classes. For example, let us consider a sample of 4 provinces and 10 sugar shops from each province. We note down the sugar prices of these 40 shops. The variation between the prices of 10 shops from each province is (heir internal variation and the totality of this interval variation is variability will) in provinces (i.e. variability with in classes).

The variation between the four sample means (i.e. average prices for each province separately) IS variability between provinces (i.e. between classes). Thus, there are two types of variations in the data one is between classes and another is within classes.

9.5 APPLICATION (OR UTILITY) OF ANALYSIS OF VARIANCE

This technique of studying homogeneity of population by separating the total variation into its various components has a much now than was conceived by Prof, R A Fisher who used it i data. Now it is applied in handling the statistics of multiple groups in various other branches of study.

(i) The main application of the analysis of variance is to test the homogeneity of the observations. In fact, it is an improvement upon student's t-test for testing the significance of several means taken together as a group The level of significance is increased in t-test because even when the sample means do not vary significantly, one or more calculated t-values may exceed from the tabulated value of 't' to make the result significant at that level. The t-test reduces precision also in estimating the population variance by the measurements of only two groups at a time. In analysis of variance all the sample means are studied simultaneously and the -population variance is estimated through the pooled variance.

Some of other applications of this technique are given below. The technique

(a) To test the significance of additional terms in regression equation.

(b) To test the curve linearity or non-linearity of the fitted regression line.

(c) To test the significance of correlation ratio.

(d) To test the significance in cases of multiple regressions.

9.6 BASIC ASSUMPTIONS

(i) The observations are independent and are distributed about a true but unknown mean μ, say.

(ii) The parent population distributions are normal with common but unknown variance a^2. Suppose variations are taken from k population (often called sub-population) having means $\mu_1, \mu_2, - - - \mu_k$ and variances $\sigma_1^2, \sigma_2^2, - - - \sigma_k^2$ respectively.

For validity of the F-test it is necessary to assume all the k populations are normal having a common variance $\sigma_1^2 = \sigma_2^2 =, - - - = \sigma_k^2$ respectively.

(i) Each observation is composed of a general (inherent) effect and the relevant effects under study are in additive form.

(ii) The errors attached to each observation are independently and normally distributed with mean zero and variance σ^2.

(iii) The variances are homogeneous. When all the sub populations have the same variances they are said to be homoscedastic.

(iv) There should be at least two observation within each class otherwise analysis of variance technique cannot be applied.

9.7 LIMITATIONS AND PRECAUTIONS IN THE ANALYSIS OF VARIANCE

(i) If the data given are in the form of proportions or in percentages then the fundamental assumptions will not appear to be fulfilled and the analysis as much cannot be done.

(ii) Sometimes various errors or factors in the field experiment may be correlated. If so, a check on them is must.

(iii) In case one or more assumptions are not satisfied in the data, the analysis can be improved by the omission of abnormal observation or treatment or the replicate as the case may be or by sub-division of the error variance or by proper choice of the scale through transformation of the data.

(ii) There should at least be two observations in each class otherwise no analysis of variance can be made.

9.8 TECHNOQUE IN ONE WAY ANALYSIS OF VARIANCE

Description and notations:

Let there be N observations classified into K classes $A_1, A_2 - - - A_k$ according to a certain factor or criterion of classification with n_i observations in A_1 classes where i= 1, 2, 3, - - - k and $\sum_{i=1}^{k} n_i = N$. Such scheme of classification according to single criterion is called one-way classification and its analysis of variance is known as one way analysis of variance.

Let X_{ij} denotes jth value in the ith class for i=1, 2, 3 - - - n_i

T_i: is the sum of the observations in A_i classes $= \sum_{j=1}^{n_i} X_{ij}$

$\overline{T}_i$: is the mean of A_i classes $= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$

T: is the sum of all the observations $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$

$\overline{X} = \overline{T} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$: overall mean

Classes	Οβσερπατιονσ	Τοτωλ	Μεανσ
A_1	$X_{11}, X_{12}, - - - X_{1j} - - - X_{1n_1}$	T_1	$\overline{T}_1$
A_2	$X_{21}, X_{22}, - - - X_{2j} - - - X_{2n_2}$	T_2	$\overline{T}_2$
–	–	–	–
–	–	–	–
–	–	–	–
A_k	$X_{i1}, X_{i2}, - - - X_{ij} - - - X_{in_k}$	T_k	$\overline{T}_i$
		T	$\overline{T}$

The following steps will be helpful in reaching to a valid inference from analysis of variance:

1. Null Hypothesis :

The null hypothesis is denoted by H_0 and is given by

$H_0 : \mu_1 = \mu_2 = \mu_3 - - - = \mu_k$

Against the alternative hypothesis and is denoted by H_1 i.e.

$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 - - - \neq \mu_k$

The model used here is

$X_{ij} = \mu + \alpha i + e_{ij}$

Where is the general effect , is the overall mean/ Population mean)

α : the additive effect and

e_{ij} : is the error effect due to chance

2. Partition of sum of squares and calculations :

(a) First of all we calculate correction factor by using the formula

$$\text{C.F.} = \frac{(\text{Grand Total})^2}{N} = \frac{T^2}{N}$$

(b) Next step is to calculate the total sum of squares by using the formula:

$$\text{T.S.S} = \sum_i \sum_j X_{ij}^2 - \text{C.F.}$$

(c) Next step is to calculate sum of squares due to treatments as

$$\text{S.S.T} = \sum_i \frac{T^2}{n_i} - \text{C.F.}$$

(d) And lastly, we have to determine error sum of square as

$$\text{SSE} = \text{TSS} - \text{SST}$$

(e) The above mentioned procedure of statistical analysis mm be represented in the following table called ANOVA Table

Sources of variation	S.S.	D.F.	M.S.S	F_{cal}
Treatment	S.S.T	K-1	S.S.T/K - 1	M.S.S.E/M.S.S
Error	S.S.E	N-K	S.S.E/N - k	

F_{tab} can be seen from F-table at (K-1) and (N-K) degrees of freedom. If the tabulated value of F is greater than F_{cal} then accept H_0 otherwise reject H_0.

9.9 TECHNIQUE IN TWO-WAY ANALYSIS OF VARIANCE

Description and Notations :

Suppose the N observations are classified into n-categories (or classes) $A_1, A_2, A_3 - - - A_n$ according to some criterion A and further into K-subclasses according to some criterion B, having nk cells or combinations

$A_i B_j$, $i = 1, 2, 3 - - - n$ and j= $1, 2, 3 - - - k$. This scheme of classification and its analysis is called two way analysis of variance.

Let X_{ij} denotes the observations in the i^{th} category of A and j^{th} category of B, $i = 1, 2, 3 - - - n$ and j

$X_{i.} = \sum_j X_{ij}$ = Sum of squares of the observations of A_i

$\bar{X}_{i.} = \frac{1}{k}\sum_j X_{ij}$ = Mean of the observations of A_i

$X_{.j} = \sum_i X_{ij}$ = Sum of squares of the observations of B_j

$\bar{X}_{j.} = \frac{1}{n}\sum_i X_{ij}$ = Mean of the observations of B_j

$X_{..} = T = Grand\ Total = \sum_i \sum_j X_{ij}$

$\bar{X}_{..} = \frac{1}{nk}\sum_i \sum_j X_{ij}$

$B_j \rightarrow$ / $A_i \downarrow$	Observations $B_1, B_2, - - - B_j - - - B_k$	Total	Means
A_1	$X_{11}, X_{12}, - - - X_{1j} - - - X_{1k}$	$X_{1.}$	$\bar{X}_{1.}$
A_2	$X_{21}, X_{22}, - - - X_{2j} - - - X_{2k}$	$X_{2.}$	$\bar{X}_{2.}$
–	–	–	–
–	–	–	–
–	–	–	–
A_i	$X_{i1}, X_{i2}, - - - X_{ij} - - - X_{ik}$	$X_{i.}$	$\bar{X}_{i.}$
–	–	–	–
–	–	–	–
–	–	–	–
A_n	$X_{n1}, X_{n2}, - - - X_{nj} - - - X_{nk}$	$X_{n.}$	$\bar{X}_{n.}$

The following steps will be helpful in reaching to a valid inference from Analysis of variance

(i) Model :

Here our model is

$$X_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

Where $\sum \alpha_i = \sum \beta_i = 0$

Here α_i. is the additional effect due to first factor, β_i. is the additional effect due to second factor, μ is the fixed effect and e_{ij} is the error term.

(ii) Null Hypothesis and Alternate hypothesis :

$$H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3} - - - = \mu_{.k}$$

H_1 : atleast two of the means $\mu_{.j}$ are not equal

$$H_0 : \mu_{1.} = \mu_{2.} = \mu_{3.} - - - = \mu_{k.}$$

H_1 : : atleast two of the means are not equal

(iii) Calculation of different sum of squares :

(a) Correction factor (C.F) i.e.

C.F.= $\frac{(X_{..})^2}{nk} = \frac{T^2}{nk}$

(b) Total sum of squares (TSS) i.e.

T.S.S = $\sum_i \sum_j X_{ij}^2 - C.F.$

(c) Sum of squares due to treatments as

S.S.T = $\sum_i X_{i.}^2 - C.F.$

(d) Sum of squares due to blocks (SSB) i.e

S.S.B = $\sum_j X_{j}^2 - C.F.$

(e) Sum of squares due to errors (SSE) i.e.,

S.S.E = T.S.S - S.S.T - S.S.B

(f) ANOVA TABLE

Sources of variations	Sum of squares	Degree of freedom	M.S.S	F_{cal}	F_{tab}
Treatment	S.S.T	n-1	S.S.T/n-1	M.S.S.T/M.S.S.E	$F_{(n-1)(k-1)}$
Blocks	S.S.B	k-1	S.S.B/k-1	M.S.S.B/M.S.S.E	
Error	S.S.E.	(n-1)(k-1)	S.S.E/(n-1)(k-1)		
	T.S.S	nk-1			

(g) Conclusion :

If F_{cal} is less than or equal to F_{tab} we accept H_0 otherwise we reject it.

9.10 SELF ASSESSMENTS

(i) Make an analysis of variance of the following data. Also, apply t-test in testing if the third and 4th class means are significantly different Classes

I	II	III	IV
7	6	8	7
2	4	4	4
4	6	5	2
	5		5

(Ans : TSS = 40.99; SSA = 3.84)

(ii) Explain the meaning and definition of the analysis of variance

(iii) State some applications of the analysis of variance.

(iv) Explain the meaning, definition, application and assumptions of analysis of variance.

(v) Explain how you will apply F-test to test the hypothesis

$$\mu_{.1} = \mu_{.2} = \mu_{.3} - - -= \mu_{.k}$$

9.11. SUMMARY

In this lesson we have introduced the concept and need of analysis of variance along with its fundamental assumptions and various possible applications. The analysis of variance based, on one way and two way classifications has also been discussed along with the simple illustration.

In the analysis part the description assumptions, model, hypothesis to be tested, least square estimates portion of various sum of squares, degree of freedom. mean sum of squares etc. have been explained in the simple manner.

DESIGN OF EXPERIMENT

STRUCTURE

10.1 Objectives

10.2 Introduction

10.3 Design of Experiment

10.4 Essentials of a Good Design

10.5 The problem of the Designing of Experiment

10.6 Basic principles of an Experimental Design/ Requirement of a good Experimental Design

10.7 Steps in Designing of an Experiment

10.8 Factors responsible for replication

10.9 Self Assessments

10.10 Summary

10. 1. OBJECTIVE

After successful completion of this lesson student will be able to :

- Understand the meaning and purpose of design of experiment.
- Know the meaning of different terms used in design of experiment
- Know the basic principles and their applications of analysis of variance

- To provide the fundamental assumptions of analysis of variance
- To introduce the concept of classification.

10.2. INTRODUCTION

Any experiment may be absolute or comparative. Absolute experiments are designed for determining the absolute values of some characteristic, like: determination of the average I.Q. of some class or the correlation coefficient between two sets of observation etc. comparative experiments are designed to compare the effect of two or more treatments on some population characteristic, like comparison of the effects of the different kinds of fertilizers on yield, etc.

Designing of an experiment means deciding how the observations or measurements should be taken to answer a particular question in a valid (strong) efficient and economical way. Thus the subject matter of design of experiment includes:

(i) Planning of the experiment.

(ii) Obtaining relevant information.

(iii) Making a statistical analysis of the data.

Experience has shown that proper consideration of the statistical analysis before the experiment is conducted, forces the experimenter to plan more carefully the design of experiment. The observations obtained from a carefully planned and well designed experiment in advance give entirely valid inferences.

10.3. DESIGN OF EXPERIMENT

Definition

Design of experiment is a logical planning (or construction) of the experiment having a complete sequence of steps taken ahead of time to ensure that the appropriate data will be obtained in a way which permits an objective analysis of a particular problem leading to valid and precise inference in most economic and useful forms.

Thus, experimental designs concern the arranging of treatments in such a manner that the inferences and conclusions regarding the effects of these treatments can easily be obtained and their reliability can duly, be measured.

The subject matter of the design of experiment includes :

(i) Planning of the experiment.

(ii) Obtaining data from it and

(iii) Making statistical analysis of the data obtained.

Before going into the details of designs of experiments, we define some important terms which are frequently used in the subject of design of experiments. Basically these terms were described in terms of agricultural experiments, but from the nature of the experiment the same designs are also used in various other situations, for example, in industry, in medical research and in educational research, etc.

(i) Treatment :

The different procedures or objects under comparison in an experiment are called treatments. For example, the different varieties of wheat or the different procedures of cultivation are treatments, in a medical experiment the different drugs are treatments, in some biological experiment the different animals are treatments.

(ii) Experimental unit or Plot :

An experimental unit is the object to which a treatment is applied in a single trial of the experiment and on which the variable under study is measured and analyzed. For example, the unit may be a plot of land, patient in hospital, an animal etc.

(iii) Experimental material or Field :

The set of the experimental units is referred to as experimental material or the experimental field. However, experimental unit is the basic unit of the experimental material.

(iv) Block :

The grouping of the experimental material in such a manner that the units within one group are alike than the units of other groups is called "Blocking or Stratification" and such groups are called blocks.

(v) Yield :

The measurement or observation obtained from an experimental unit is called Yield.

(vi) Precision :

The degree of certainty with which the influences are drawn from the result of experiment is called precision. The magnitude of the degree of uncertainty is assessed by the reciprocal of the variance of mean. This is the ability of an experiment with which it detects the assignable differences between the treatments. This is also called "sensitivity" or "amount of information."

(vii) Accuracy:

The accuracy of a design is a measure for the absence of bias and has been used as a synonym for precision. The lesser the bias, the greater is the accuracy of the design.

(viii) Experimental error :

It is a characteristic of the experimental units that they produce different results on repetitions. Not only the different units but the same unit give different results at different timings and in different hands owing to the human variation in taking observations, or owing to instrument variation in indicating the effect, measurement. There are many sources of these variations. A fundamental phenomenon in various experiments is that some of the variations are systematic and may be detected but some are apparently random and cannot be explained. The variation due to uncontrolled factors is named "Experimental error."

The main purpose of the experimentation is to compare the different treatments by separating the systematic and the experimental variations or error.

10.4. ESSENTIALS OF A GOOD DESIGN

A design means the specifications of

(i) The treatments to be compared,

(ii) The plots or the experimental units to which the treatments are to be applied.

(iii) The method to be followed for allocation of the different treatments to the plot and

(iv) The procedure of analysis, which of course will depend on the method of the allocation of treatments to the plots.

To have a good design, we need following considerations :

(i) The designs of experiment should be as simple as possible and it should have sufficient accuracy to accomplish it purpose.

(ii) The result of the experiment should have sufficient scope. The scope and the form of the layout is decided by the objective of the experiment which is indirectly defined by the null hypothesis.

(iii) There should not be any bias in the experiment. In other words, the design should be uniform w.r.t the various treatments without favouring either of them.

(iv) The design should be able to give a valid measure of the experimental error, by making a sufficient control over undesirable influences i.e, the inferences drawn should be valid and precise.

(v) The degree of uncertainty with which the inferences are drawn may be well defined and it should be possible to device suitable unbiased measures of the treatments effect.

(vi) It should be possible to s the magnitude of degree of uncertainty from the results of the experiment by the use of the statistical theories based on probabilities.

10.5. THE PROBLEMS OF DESIGNING OF EXPERIMENT

If we go on repeating some treatments, it would be found that its effect varies from trial to trial; so much so that even after a very large number of trials, the investigator is not absolutely certain that how his result would turn out if some more trials were given to the same experiment and under similar conditions.

The Problems of the design of experiment consists of estimating the actual difference between the average yield of several treatments and making their statistical analysis. It is for the various types of improvements, accuracy and other details that the different type of designs have been invented. The following are the main considerations in the planning of an experiment

i) Statement of the problem and its objective.

ii) Formulation of the hypothesis

iii) Description of the experiment

iv) Performance and summarization of the experiment.

v) Statistical analysis and test of significance,

vi) Drawing conclusions,

i) Preparation of the reports.

10.6. BASIC PRINCIPLES OF AN EXPERIMENTAL DESIGN/ REQUIREMENT OF A GOOD EXPERIMENTAL DESIGN.

Designing of an experiment stands for deciding how; the valid and efficient treatment pervasions or measurements can be taken in the most economic way. The study of the experimental design was pioneered by Prof. R. A. Fisher. According to him there are three basic principles of the design of experiments; namely

i) Replication

ii) Randomization and

iii) Local control.

The first two i.e. replication and randomization are indispensable for an experimental design and the third one i.e. the local control remains desirable.

i) Replication :

A treatment is repeated a number of times in order to obtain more reliable estimate and that is not possible from a single observation. The repetition of the treatments under consideration is called replication. If we want to compare the effect of some treatments we will have to allot them to a number of plots. The number of plots to which a treatment is allocated is known as replication.

ii) Randomization :

The allocation of treatments to various plots (or experimental units) at random is called randomization. A set of treatments applied to a set of experimental units is said to be randomized if the treatment applies to any given unit is chosen at random from the available treatments and not already allocated. The process of randomization consists in providing

equal chance to all plots to be put under any particular treatment. This is done with the help of random number table. In brief; the allocation of treatments to various plots at random is called randomization; the main objectives of the randomization are :

(i) Randomization give equal chance or scope to various treatments to show their merits and to ensure that on the average that different treatments are subject to some environmental conditions.

(ii) It is a device to eliminate bias.

(iii) The validity of the test of significance in the ANOVA is based on the assumption of randomization.

(iii) Local Control / error Control :

The method that are adopted to reduce the experimental error without increasing unduly the replications or without interfering with the statistical requirements of randomness or without sacrificing average are known as the technique for local control or error control.

The most important technique of error control is often called local control is a technique that handles the experimental material in such a way that the effects of variability are reduced. This reduction depends on the size, shape and structure of experimental units and on the division of basic units into relatively homogenous groups in brief. If the experimental material says field for agricultural experimentation is heterogeneous and different treatments are allocated to various units (plots) at random over the entire field, the soil heterogeneity will also enter the uncontrolled factors and thus increase the experimental error as for as possible In addition to the principles of replication and randomization the experimental error further be reduced by making use of the fact that neighboring areas in a field are relatively more homogenous than those widely spread. In order to reduce the experimental error the whole experimental area (field) is divided into homogenous groups (blocks) row wise or column wise or I both in such a manner that the variation within each block is minimum. The process of reducing the experimental error by dividing the relatively heterogeneous experimental area into homogenous blocks is called as local control or error control.

10.7. STEPS IN DESIGNING OF AN EXPERIMENT

Following are the main steps in designing of an experiment :

i) Statement of the problem or the object :

In absence of a clearly defined object, we can not experiment the data nor can the exact scope of the experiment be decided.

ii) Scope :

Every design has its own scope and gives valid conclusions within its scope only. However, it should possibly be fetched to the maximum.

iii) Study of experimental material :

To make the design efficient, it should be as homogenous as possible.

iv) Choice of the experimental technique :

The choice of the design depends upon the heterogeneity of the experimental material, number and nature of treatments and upon the desired precession

v) Performance of the experiment :

Every design has its own way of performance, It may include the problem of the size and shape of the plots and blocks also.

vi) Collection of the data and their statistical analysis :

The data collected and arranged systematically for statistical analysis which comprises of analysis of variance.

vii) Conclusion :

It is drawn from the statistical analysis and is given in the form of a report.

10.8. FACTORS RESPONSIBLE FOR REPLICATION

There are mainly nine factors which are responsible for determining the number of replications.

i) Extent of precision required.

ii) Heterogeneity of experimental material.

iii) Availability of resources.

iv) Size of experimental unit.

v) Required degree of freedom of experimental error.

vi) The relative cost of experimental error.

vii) The extent and nature of competition among experimental units.

viii) The number and nature of the treatments.

ix) The fraction to be sampled.

10.9. SELF ASSESSMENTS

i) Calculate the minimum number of replications required so that an observed difference of 10% of the mean will be regarded as significant at 5% level of significance. The coefficient of variation of plot values being 12%.

ii) Calculate the minimum number- of replications so that an observed difference of 5% of the mean will be taken as significant at 1% level of the significance. The coefficient of variation of plots values being 15%;

10.10. SUMMARY

In the present lesson we have discussed the concept of design of experiment and various terms used in any design. The terms randomization, replication and local control also been discussed in detail in this lesson.

B.A. SEMESTER IV **UNIT - III**

ST - 404 **LESSON NO. 11**

COMPLETE RANDOMIZED DESIGN

STRUCTURE

11.1 Objectives

11.2 Introduction

11.3 Definition

11.4 Some Important Terms Used in Design of Experiment

11.5 Complete Randomized Design

11.6 Simple Illustrations

11.7 Self Assessment

11.8 Summary

11.1 OBJECTIVES

After successful completion of this lesson students will be able to

- Understand the meaning and purpose of design of experiment.
- Know the meaning of different terms used in design of experiment.
- Know the completely ^Randomized design and its analysis part.
- Perform analysis of completely Randomized design (C.R.D)

11.2. INTRODUCTION

Experimentation and making inferences are twin essential features of general scientific methodology. Statistics as a scientific discipline is mainly designed to achieve these objectives. It is generally concerned with problems of inductive inferences in relation to stochastic models describing random phenomena. When faced with the problems of studying a random phenomenon, the scientist, in general, may not have complete knowledge of the true variant of the phenomenon under study. A statistical problem arises when he is interested in the specific behavior of the unknown variate of the phenomenon. After a statistical problem has been set up, the next step is to perform experiments for collecting information on the basis of which inferences can be made in the best possible manner.

The methodology for making inferences has three main aspects. First, it derives methods for drawing inference from observations when these are exact but subject to variation. As such, the inferences are not exact but probabilistic in nature. Second it specify methods for collection of data approximately so that the assumptions for the application of appropriate statistical methods to them are satisfied. Lastly the techniques for proper interpretation of results are derived. A good deal of work has also been done in the field of data collection and interpretation techniques. The topic of the lesson discussion is design and analysis of experiments which falls in the sphere of data collection and interpretation techniques. The other main topic in this regard is the theory of sample surveys. Though the theories of sample surveys and design of experiments are both concerned with data collection techniques, they serve different purpose. The theory of sample surveys has the objective of observations from a population which exists in its own way such that the sample can adequately represent and accurately interpret the population, In the case of experimental data no such population exists in its own way, What exists is a problem and the data have, so to say, to be manufactured by proper experimentation so that an answer to the problem can be inferred from the data.

11.3 DEFINITION

Researcher/scientists always plan the experiment before it is conducted. To achieve objectives relevant to the problem under investigation we need to construct the

complete sequence of steps which is called designing of the experiment. Thus design of experiment is a logical planning of the experiment having a complete sequence of steps taken ahead of time to ensure that the appropriate data will be obtained in a way which permits an objective analysis of the particular problems leading to valid and precise inference. The subject matter of the design of experiment includes: Planning of the experiment, obtaining the data and statistical analysis of the data obtained.

11.4 SOME IMPORTANT TERMS USED IN DESIGN OF EXPERIMENT

Before going into the detail of design of experiment, we first define some important terms which are used frequently.

Treatment:

Different objects under comparison in an experiment are called a treatment that is by a treatment we mean a particular set of experimental conditions which are imposed on experimental unit within the confines of chosen design. For example, in agronomy a treatment might refer to brand of fertilizer or amount of fertilizer.

Replication:

The repetition of treatment or treatment combination under study is called replication. If a single treatment is applied r times in an experiment we say that the treatment has been replicated r times.

Experimental Unit:

The unit to which a single or combination of treatments is applied in one replication of the experiment is called the experimental unit or unit or plot.

Experimental Error:

The failure of two identically treated experimental unit to yield identical results is termed as experimental error. In a particular experiment the experimental error reflects

(i) Errors of observation

(ii) Errors of measurement

(iii) Errors of experimentation

(iv) Variation of the experimental material

(v) The combined effects of all uncontrollable factors which can influence the characteristics under study.

The experimental error can be reduced by the following techniques

(i) By using more homogeneous experimental material

(ii) By taking more care in conducting the experiment

(iii) By using the information provided by related variates.

(iv) By using an appropriate experimental design. Use of appropriate design will effectively reduce the error.

Lay out:

The term lay out means the allocation of treatments to plot in accordance with the conditions of the design.

Effect and Interaction:

The effect of a factor is a measure of the change in the response variable to the changes in the levels of the factor averaged over all the levels of the other factors. Interaction is an additional effect which occurs due to the combined influence of two or more factors.

Models and Analysis of Variance:

A statistical model is linear relation of the effects of the different levels of a number of factors involved in an experiment along with one or more terms representing error effects. The effects of any factor may either, be fixed or random. For example, the effects of two well defined levels of irrigation are fixed as each irrigation level can be reasonably taken to have a fixed effect.

In this lesson, we shall restrict ourselves to mainly fixed effect model. Further we also assume in the models that the effects are additive. After a model has been fixed, the general method of analysis of variance is used to analysis the data.

11.5 COMPLETELY RANDOMIZED DESIGN (C.R.D)

Introduction:

It is the simplest type of design in which the basic assumption is that the entire experimental material is homogeneous and the experimental units are assigned completely at random to the different treatments or vice versa. Thus every experimental unit has the equal chance of receiving a particular treatment. Allocation of the units is done with the help of random number table, hi this design the principle of replications and randomization is used.

Description:

Let there be v treatments in all and the i^{th} treatment be replicated r_1. times i = 1, 2, - - - v. The total number of plots required to lay of the experiment will be n = $r_1 + r_2 + r_v$. In a CRD each of the v treatments is applied randomly such that i^{th} treatment occurs in r. plots, each plot being selected at random.

Lay out:

In a CRD we divide the experimental material into n plots of equal size and randomly apply the v treatments so that i^{th} treatment occurs exactly r. times.

Statistical Model:

The observations recorded in CRD can be expressed in the form of the following model:

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

where y_{ij} is the observation obtained from the jth plot to which i^m treatments has been applied, μ represents a general effect (Mean), α_i. is the effect of the ith treatment and e_{ij} is the error associated with the ith treatment and j^{th} plot which is a random variable and are independent N(0, σ^2) random variables.

Assumptions:

The following are the main assumptions which are required to use the analysis of variance technique to analyse the data.

(i) The treatment effects are additive and

(ii) The errors are independent and normally distributed with mean zero and constant variance σ^2

Formulation of Hypothesis:

In CRD we formulate the null and alternate hypothesis as

$H_0: \alpha_1 = \alpha_2 = \text{- - -} = \alpha_v$ i.e. there is no significant differenc en the effects of the different treatments and the alternative hypothesis is

H_1 : At least two of the treatment effects are not equal

Analysis of the Data: To analyse the CRD, the following are the main steps:

Step 1: First of all complete the following table from the given data.

Treatments	Observations	Total	No of observations	Means
T_1	$y_{11}, y_{12}, - - - y_{1j} - - - X_{1r_1}$	T_1 .	r_1	$\bar{y}_1$.
T_2	$y_{21}, y_{22}, - - - y_{2j} - - - y_{1r_2}$	T_2 .	r_2	$\bar{y}_2$.
–	–	–	–	–
–	–	–	–	–
–	–	–	–	–
T_i	$y_{i1}, y_{i2}, - - - y_{ij} - - - y_{1r_i}$	T_i .	r_i	$\bar{y}_i$.
–	–	–	–	–
–	–	–	–	–
–	–	–	–	–
T_v	$y_{v1}, y_{v2}, - - - y_{vj} - - - y_{vr}$	T_v .	r_v	$\bar{y}_v$.
Overall Total mean		T	n	$\bar{y}_{..}$

Step II:

Determine the correction Factor which is usually denoted by C.F as

$$\text{C.F.} = \frac{T^2}{n}$$

Step III: Find the total sum of squares (TSS) on using

$$\text{TSS} = \sum_i \sum_j \left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_i \sum_j y^2{}_{ij} - \frac{T^2}{n} = \sum_i \sum_j y^2{}_{ij} - C.F.$$

i.e TSS = Sum of Squares of Observations - Correction Factor

Step IV: Find the sum of squares due to treatments which is denoted by SST as

$$\text{SST} = \sum_i^v r_i \left(\bar{y}_i - \bar{y}_{..}\right)^2 = \sum_i^v \left(\frac{T_{i.}^2}{r_i}\right) - \frac{T^2}{n} = \sum_i^v \left(\frac{T_{i.}^2}{r_i}\right) - C.F.$$

Step V: Find the sum of squares due to Error (residual sum of squares) which is denoted by RSS or SSE as

$$\text{SSE} = \sum_i \sum_j \left(y_{ij} - \bar{y}\right)^2 = \sum_{i=1} \sum_j y_{ij} - \frac{T_i^2}{n_i}$$

Note: You may verify that T.S.S = S.S.T + S.S.E or S.S.E = T.S.S - S.S.T

Step VI: Find the degree of freedom of different components as

D.f. for TSS = n - 1

D.f. for SST = v - 1

D.f. for SSE = n - v

Step VII: Obtain the variance between the treatments and within the treatments by dividing the sum of squares of each by its corresponding degree of freedom, which is also called Mean Sum of Squares.

Step VIII: Compute F-statistics as

F = (variance between treatments)/ (Variance within the Treatments)

Step IX: Compare the value of F statistic with tabulated value of F at the desired level of significance (a) with v- 1 and n - v degree of freedom. If calculated value of F is greater than the tabulated value, we reject the null hypothesis, otherwise accept the null hypothesis.

It is customary to summaries the above steps in the form of following table which is called Analysis of Variance table (ANOVA TABLE)

ANOVA TABLE

Sources of variations	Sum of squares	Degree of freedom	M.S.S	F_{cal}	F_{tab}
Treatment	S.S.T	v -1	S.S.T/v -1	$\frac{S.S.T\ (n-v)}{S.S.E\ (v-1)}$	$F_{(v-1)(n-v)}$
Error	S.S.E.	(n-v)	S.S.E/(n-v)		
	T.S.S	n-1			

Note: In case H_0 is rejected, we further examine as to which pair of treatments is significantly different with the help of student's t-test. To compare i^{th} and j^{th} treatment we calculate the value of Critical Difference (CD)

$$C\ D = t_{n-v}(\alpha)\sqrt{MSE(\frac{1}{r_i}+\frac{1}{r_j})}$$

If the difference in no treatment means in questions is greater than CD then they are said to differ significantly at the given level of significance.

Applications:

This design is applied in the laboratory experiments and in green houses experiments where one can obtain the homogeneous material for plots. The C.R.D is recommended in situations where an appreciate fraction of the experimental units are likely to be destroyed or fail to respond.

11.6. SIMPLE ILLUSTRATIONS:

Example 1:

Four groups of pigs were given different feeds F_1, F_2, F_3 and F_4 and their increase in weights (in kg) were recorded by an experiment are given below. Do the feeds differ each other as far as the phenomenon of increasing of weights is concerned?

Feeds	Increase in weights (in kg)				Total	Mean
F_1	8	10	14	7	39	9.75
F_2	12	15	14	13	54	13.5
F_3	6	8	6	9	29	7.25
F_4 .	10	7	9	-	26	8.65

Solution :

First of all we formulate the null and alternative hypothesis as

H_0: There is no significance difference between the performances of feeds.

H : The performance of at least two feeds differs significantly.

To test the null hypothesis we use the procedure and analysis of C.R.D and first obtain the following sum of squares as:

Correction Factor (C.F) $= \frac{T^2}{n} = \frac{148^2}{15} = 1460.267$

Total sum of Squares (T.S.S) = $\sum y_{ij}^{\,2} - C.F.$

$= 8^2 + 10^2 + ... + 7^2 + 9^2 = 1460.267 = 1590 = 1460.267 = 129.733$

Sum of Squares due to feeds (S.S.T) $= \sum \frac{T_i^2}{r_i} - C.F.$

$$= \frac{39^2}{4} + \frac{54^2}{4} + \frac{29^2}{4} + \frac{26^2}{4} - 1460.267$$

$= 1544.833 - 1460.267 = 84.566$

Sum of squares due to Error (S.S.E) = T.S.S - S.S.T

$= 129.733 - 84.566 = 45.167$

Now we summaries the above calculation in the following ANOVA Table:

Sources of variation	S.S	d.f.	M.S.S	F-ratio
Feeds Error	84.566 45.167	3 11	28.189 3 4.166	6.87
Total	129.733	14		

If we take level of significance 5% then the table value of F-statistic with 3 and 11 degree of freedom at 5% level is 3.59. Since F-ratio (calculated value of F) is greater than the tabulated value of F, hence we reject the null hypothesis and conclude that at least two feeds differ from each other. To test the difference in means of any two feed, we calculate critical difference (CD). Suppose we want to compare feed F_2 and F_4 then value of CD is

$$\text{C.D.} = t_{11}(0.05)\sqrt{MSE\left(\frac{1}{r_i}+\frac{1}{r_j}\right)} = 2.201\sqrt{4.106\left(\frac{1}{4}+\frac{1}{3}\right)} = 3.41$$

The difference in means of F, and R is 13.5 - 8.67 = 4.83, since this is greater than the value of CD. So we conclude that the feed F_2 and F_4 differ significantly from each other.

11.7.SUMMARY

In the present lesson we have discussed the concept of design of experiment and various terms used in any design. The basic principles of design of experiment namely-Replication, Randomization and Local Control have also been explained in detail. The layout and analysis of completely randomized design, which depends on only replication and randomization, have also been discussed with simple illustration,

LL.8. SELF ASSESSMENTS:

Q.l. What do you understand by design of Experiment?

Q.2. Define the following terms:

(i) Experimental Material

(ii) Treatment

(iii) Experimental Unit and

(iv) Experimental Error

Q.3. Explain the basic principles of design of experiment.

Q.4. Give the layout and analysis of C.R.D underlying its various assumptions.

SAMPLING THEORY
RANDOMIZED BLOCK DESIGN

STRUCTURE

12.1 Introduction

12.2 Randomized Block Design

12.3 Advantages and Disadvantages of R.B.D

12.4 Application of R.B.D

12.5 Simple Illustrations

12.6 Summary

12.7 Self Assessments

12.1. INTRODUCTION

The Randomized Block Design (RBD) is the simplest design which uses all the three principles enunciated by Fisher. The one of the basic assumption of Completely Randomized Design was that the experimental material must be homogeneous, that is, in case of heterogeneity of experimental material one cannot use completely randomized design. To overcome this difficulty we use the third principle of design-local control. As explained in previous lesson, the local control is a technique that handless the experimental material in such a way that the effect of variability is reduced. The design which employed all the three principles is randomized block design.

12.2. RANDOMIZED BLOCK DESIGN

The C.R.D is not useful if the experimental units are not homogeneous as the variation among the unit will vitiate the test of significance of the treatment effects. The simplest design which enable us to takes care of the variability among the units is R.B.D. This design is based on all three principles of design of experiment. Suppose we want to compare the effects of 't' treatment, each treatment being replicated an equal number of times say r times. Then we need n = rt experimental unit and these units are not perhaps homogeneous. The R.B.D consists of two steps. The first step is to divide the units into r more or less homogeneous groups or block. In each blocks, we take as many unit as there are number of treatments. Thus number of blocks is equal to the replication of each treatment (r). Then treatments are assigned at random to the experimental units within each block and this is done independently for each block.

Layout:

If there are t treatments, then the r (number of replications of each treatment) blocks are formed in a direction perpendicular to their fertility gradient such that each block contains relatively homogeneous material. Each block is further divided into t plots of equal size. Within each block treatments are randomly allocated.

Model and Assumptions:

The observations recorded in R.B.D can be expressed in the form of following model:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

Where y_{ij} is the observation obtained from j^{th} block to which i^{th} treatment has been applied. μ is the general effect, α_i is the effect of the i^{th} treatment, β_j is the effect of jth block and e_{ij} (random component) are independently and normally distributed with mean zero and constant variance. In this model treatment

and $\sum \alpha_i = \sum \beta_j = 0$

Here the null hypothesis is under test are:

$$H_0 : \alpha_1 = \alpha_2 = - - - = \alpha_t$$

$$H_0 : \beta_1 = \beta_2 = - - - = \beta_r$$

And alternate hypothesis are:

(i) H_1: At least two treatment effects differ significantly.

(ii) H_1: At least two block effects differ significantly.

The analysis of data can be done by using two way analysis of variance analyses as with t treatment r replications (blocks) the observation in R.B.D can be arranged as follows:

Treatments	Blocks B_1 B_2 --- B_j ... Br	Total	Means
T_1	$y_{11}, y_{12}, - - - y_{1j} - - - X_{1r_1}$	T_1.	$\bar{y}_1$.
T_2	$y_{21}, y_{22}, - - - y_{2j} - - - y_{1r_2}$	T_2.	$\bar{y}_2$.
–	–	–	–
–	–	–	–
–	–	–	–
T_i	$y_{i1}, y_{i2}, - - - y_{ij} - - - y_{1r_i}$	T_i.	$\bar{y}_i$.
–	–	–	–
–	–	–	–
–	–	–	–
T_t	$y_{v1}, y_{v2}, - - - y_{vj} - - - y_{vr}$	T_t.	$\bar{y}_v$.
		T	$\bar{y}$..

After completion of above table, we find different sum of squares:

$$\text{TSS} = \sum_i \sum_j y_{ij}^2 - \frac{T^2}{rt}$$

$$\text{SST} = \sum_i \left(\frac{T_{i.}^2}{r}\right) - \frac{T^2}{rt}$$

$$\text{SSB} = \sum_j \left(\frac{T_{.j}^2}{t}\right) - \frac{T^2}{rt}$$

$$\text{SSE} = TSS - SST - SSB$$

In R.B.D the analysis of variance (ANOVA) table will be as follows:

ANOVA TABLE

Sources of variations	Sum of squares	Degree of freedom	M.S.S	F_{cal}	F_{tab}
Treatment	S.S.T S_T^2	(t -1)	S.S.T/t-1 $= S_T^2$	$\frac{MST}{MSE} = \frac{S_T^2}{S_E^2}$	$F_{(t-1),(t-1)(r-1)}$
Block	S.S.B. S_B^2	(r-1)	S.S.B./ (r-1) $= S_B^2$	$\frac{MSB}{MSE} = \frac{S_B^2}{S_E^2}$	
Error	S.S.E. S_E^2	(r-1)(t-1)	S.S.E./ (r-1)(t-1)$= S_E^2$		
	T.S.S	rt-1			

For testing the significance of treatments effects, we reject the null hypothesis at a level of significance if

$$F_1 = \frac{MST}{MSE} > F_{(t-1),(t-1)(r-1)}$$

Otherwise accept the null hypothesis.

Similarly, for significance of block effects, we reject the null hypothesis at a level of significance if

$$F_2 = \frac{MSB}{MSE} \quad F_{(t-1),(t-1)(r-1)}$$

Otherwise we accept the null hypothesis.

Note: If the null hypothesis for treatment effects is rejected then to test the difference in two treatments means we can use the critical difference (CD) as

$$C\,D = t_{(r-1)(t-1)}(\alpha)\sqrt{\frac{2MSE}{r}}$$

Where $t_{(r-1)(t-1)}(\alpha)$ is tabulated value of 't' statistic with (r-l)(t-l) degree of freedom at a level of significance. If the difference between two treatments mean in question is greater than the CD then they are said to be differ significantly at given level of significance.

Further, if the effect of blocks is non-significant then the blocks are pooled with error and the ANOVA table is similar to that of C.R.D.

12.3. ADVANTAGES AND DISADVANTAGES OF R.B.D

Advantages of RBD:

The following are the main advantages of R.B.D.

(i) It is an improvement over C.R.D because by making groups the variation between the blocks are eliminated from that of error variance and thus experimental error is reduced.

(ii) This design is flexible for any number of treatments and any number of replication may be used with restriction that the number of replication is equal to the number of blocks.

(iii) The analysis is simple and even with missing plots it is not much complicated.

(iv) It provides a method of eliminating or reducing the long term effects.

Disadvantages of R.B.D:

The main disadvantages of R.B.D are:

(i) Here we assume that the relative effects of treatments are the same in all the blocks. This is true when there is no interaction between blocks and the treatment jji practice, we see that the interactions are there but their effects are not appreciable.

(ii) With missing values partitioning of the treatments sum of squares becomes impossible because of non-orthogonality of the data.

(iii) If the number of treatments is very large, then the size of each group or block will increase whereas the number of blocks will decrease. This may introduce heterogeneity within blocks.

(iv) If cannot control two sides' variations of the experimental material simultaneously.

(v) We cannot use different number of replications for different treatments.

12.4. APPLICATIONS OF R.B.D

The following are the main applications of R.B.D.

(i) It is applicable to moderate number of replications.

(ii) In agricultural experiments this design is used if the variations in the soil fertility gradient are in one direction only.

(iii) This design can be used where it is desired to control one source of variation besides the variation in treatment effects.

(iv) This design is considered the backbone of science of experimental design owing to the presence of validity, simplicity and flexibility.

12.5. SIMPLE ILLUSTRATIONS

Example 1:

The data given below represents the yield of cowpeas obtained from each plots of a randomized block experiment involving four varieties of cowpeas Analyse the data and give your comments.

COWPEAS (VARIETIES)

		V_1	V_2	V_3	V_4	**Total**	**Mean**
	I	810	780	840	740	3170	792.50
	II	905	825	900	780	3410	852.50
Blocks	III	805	740	870	800	3215	803.75
	IV	925	850	810	720	3305	826.25
	Total	**3445**	**3195**	**3420**	**3040**	**13100**	**818.75**
	Mean	**760**	**818.75**	**861.25**	**798.75**	**855**	

Solution:

The above given data is of R.B.D and can be analyzed by the procedure of analysis of two way classification. First of all, we formulate the null and alternative hypothesis and calculate the different sum of squares.

H_0: The varieties of cowpeas do not differ from each other with regard to yields.

H_1 : At least two varieties differ significantly from each other Hypothesis of Blocks:

H_0: The blocks do not differ from each other with regards to yields of cowpeas.

H_1 : At least two blocks differ significantly from each other.

Sources of variation	Sum of Squares	d.f	Mean Sum of Squares	F-ratio
Verities	27887.50	3	9295.83	4.42
Block	8437.50	3	2812.50	1.34
Error	18950.00	9	2105.56	
Total	55275.00	15		

Correction Factor(C.F.) $= \frac{T^2}{rt} = \frac{13100^2}{16} = 10725625.00$

Total Sum of Squares due to varieties

(SST) $= \frac{3445^2+3195^2+3420^2+3040^2}{4} - CF = 27887.50$

Total Sum of Squares (TSS) = 1078900-10725625 = 55275

Sum of Squares due to blocks (SSB) $= \frac{3170^2+3410^2+3215^2+3305^2}{4} - CF$

= 102340602.50 – 10725625 = 8437.50

Sum of squares due to error (SSE) = 55275 – 27887.50 – 8437.50 = 18950

Since the calculated value of F ratio for varieties (4.42) is greater than the tabulated values of F statistic (3.86) with 3 d.f. at 5% level of significance, so we reject our null hypothesis of varieties and the calculated nature of F-ratio (1.34) for blocks is less than the tabulated value of F statistic (3.86) with 3 and 9 d.f. at 5% level. So we accept the null hypothesis for blocks. That is we may conclude that there is no significance difference between blocks.

As the null hypothesis of varieties was rejected, thus to test the difference in means of any varieties we can use critical difference (CD). Suppose we want to compare variety V_1 and V_4 then the value of CD is

$$\text{C.D.} = t_9(0,05)\sqrt{MSE\left(\frac{1}{r}+\frac{1}{r}\right)} = 69.11$$

The difference in means of variety V_1 and V_4 is 861.25 -760.40 = 101.25 Since this value is greater than the value of CD (69.11). So we conclude that the variety V_1 and V_4 differ significantly from each other. Similarly, we can take pairs of other varieties.

Note: In this example, Blocks are found to be non-significant, so we can pool them with error and new ANOVA can be made.

12.6. SUMMARY

For testing the significance of treatment effects, we reject the null hypothesis at α% level of significant if

$$F_1 = \frac{MST}{MSE} > F_{(t-1),(t-1)(r-1)}$$

Otherwise accept the null hypothesis.

Similarly, for significance of block effect, we reject the null hypothesis at a level of significance if

$$F_2 = \frac{MSB}{MSE} \quad F_{(t-1),(t-1)(r-1)}$$

Otherwise we accept the null hypothesis.

If the null hypothesis for treatment effect is rejected then to test the difference in two treatments means we can see the critical difference (CD) as

$$C\,D = t_{(r-1)(t-1)}(\alpha)\sqrt{\frac{2MSE}{r}}$$

where $t_{(r-1)(t-1)}(\alpha)$ is the tabulated value of 't' statistics with (r-1)(t-l) degree of freedom

12.7 SELF ASSESSMENT

(a) Explain clearly the difference between C.R.D and R.B.D. Also give the ANOVA table for both.

(b) Discuss the advantages and disadvantages of R.B.D.

(c) Give the layout and analysis of R.B.D.

(d) Analysis the following data and comment on your findings:

Plots / Blocks	A	B	C	D
1	300	330	360	29020
2	240	230	350	240
3	370	360	390	310
4	270	320	320	260

LATIN SQUARE DESIGN

STRUCTURE

13.1 Introduction

13.2 Objectives

13.3 Latin Square Design(L.S.D)

13.4 Statistical Analysis of $m \times m$ L.S.D

13.5 Estimation of missing value in L.S.D.

13.6 Illustration

13.7 Summary

13.8 Self Assessment Questions

13.1 INTRODUCTION

In this lesson we make a comprehensive study of latin Square Design viz, its meaning, Statistical Analysis, etc. Illustration has also been provided to show its practical application.

13.2 OBJECTIVES

- To introduce latin square design
- To explain statistical analysis of LSD
- To estimate the missing value in LSD

13.3 LATIN SQUARE DESIGN (L.S.D.)

In field experimentation, it may happen that experimental area (field) exhibits fertility in strips, In this case R.B.D. will be effective if the blocks happen to be parallel to these strips and would be extremely inefficient if the blocks are across the strips. Initially, fertility gradient is seldom known. A useful method of eliminating fertility variations consists in an experimental layout which will control variation in two perpendicular directions. Such a layout is a Latin Square Design (L.S.D.)

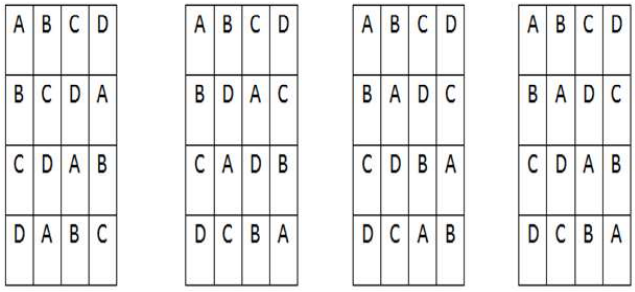
In a Latin square every row and every column is a complete replication. The Latin square provides more opportunity than randomized blocks for the reduction of errors by skillful planning. The experimental material should be arranged and the experiment conducted so that the differences among rows and columns represent major sources of variation,

For m treatments, there have to be $m \times m = m^2$ experimental units. The whole of experimental area is divided into m^2 experimental units (plots) arranged in a square so that each row as well as each column contains m units (plots). The m treatments are then allocated at random to these rows and columns in such a way that every treatment occurs once and only once in each row and in each column. Such a layout is known as $m \times m$ Latin Square Design (L.S.D.) and is extensively used in agricultural experiments.

For example, if we are interested in studying the effects of m types of fertilizers on the yield of a certain variety of wheat, it is customary to conduct the experiments on a square field with m^2-plots of equal area and to associate treatments with different fertilizers and row and column effects with variations in fertility of soil. Obviously, there can be many arrangements for an L.S.D. and a particular layout in an experiment must be determined randomly.

Standard Latin Square: A Latin square in which the treatments, say, A, B, C ... etc. occur in the first row and the first column in alphabetical order is called a St*andard Latin Square or a Latin Square in Canonical Form.*

For a 4 x 4 Latin Square Design, 4 standard squares are possible as given in Figures



1. With two-way grouping or stratification L.S.D. controls more of the variation than C.R.D. or R.B.D.

Thus, in field experimentation if the fertility gradient is in two directions at right angles to each or in one unknown direction then L.S.D. is likely to be more efficient than R.B.D. In fact L.S.D. can be used with advantage of those cases where the variation in experimental material is from two orthogonal sources. As regards the applications of L.S.D., Professor Fisher says, "If experimentations were only concerned with the comparison of four to eight treatments or varieties, it (L.S.D.) would be not merely the principal but almost the universal design employed

2. L.S.D. is an incomplete 3-way layout. Its advantage over the complete 3-way layout is that instead of m^3 experimental units only m^2 units are needed. Thus a 4 x 4 L.S.D. results in saving of $m^3 - m^2 = 4^3 - 4^2 = 64 - 16 = 48$ observations over a complete 3-way layout

3. The statistical analysis is simple though slightly complicated than for R.B.D. Even with 1 or 2 missing observations the analysis remains relatively simple.

4. More than one factor can be investigated simultaneously and with fewer trials than more complicated designs.

Disadvantages of L.S.D.

1. The fundamental assumption that there is no interaction between the three factors of variation (i.e., the factors act independently) may not be true in general.

2. Unlike R.B.D., in L.S.D. the number of treatments is restricted to the number of replications and this limits its field of application. L.S.D. is suitable for the number of treatments between 5 and 10 and for more than 10 to 12 treatments the design is seldom used since in that case the square becomes too large and does not remain homogeneous.

3. In case of missing plots, when several units are missing the statistical analysis becomes quite complex. then in R,B.D. where we can easily omit the data for these blocks without complicating the analysis at all whereas this is not possible in LSD because in LSD the number of rows, columns and treatments have to be equal.

4. In the field layout, R.B.D. is much easy to manage than L.S.D., since the former can be performed equally well on a square or rectangular field or a field of any shape whereas for the latter approximately a square field is necessary.

13.4 STATISTICAL ANALYSIS OF $m \times m$ L.S.D. for One Observation per **Experimental Unit**: Let y_{ijk} $(i, j, k = 1,2, \ldots, m)$ denotes the response from the unit (plot, in field experimentation) in the ith row, jth column and receiving the kth treatment. The triple (i,j, k) assumes only m^2 of the possible m^3 values . If S represents the set of m^2 values, then symbolically $(i, j, k) \in S$. If a single observation is made per experimental unit, then the linear additive model is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk}, \quad (i,j,k) \in S$$

where μ is the constant mean effect; α_i, β_j *and* γ_k are the constant effects due to the ith row, jth column and kth treatment respectively and e_{ijk} is error effect due to random component assumed to be normally distributed with mean zero and variance σ_e^2 i.e., $e_{ijk} \underset{\sim}{} N(0, \sigma_e^2)$. If we write

G = y... = Total of all the m^2 observations

$R_i = y_{i..}$ = Total of the m observations in the ith row

$C_j = y_{.j.}$ Total of them observations in the jth column

$T_k = y_{..k}$ = Total of the m observations from kth treatment, Then we have

$$\sum_{i,j,k\in S} (y_{ijk} - \bar{y}_{...})^2$$

$$= \sum_{i,j,k\in S} [(\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{..k} - \bar{y}_{...}) + (y_{ijk} - \bar{y}_{i..} - \bar{y}_{i..} - \bar{y}_{i..} + 2\bar{y}_{...})]^2$$

$$= m\sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 + m\sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2 + m\sum_k (\bar{y}_{..k} - \bar{y}_{...})^2 + \sum_{i,j,k\in S} (y_{ijk} - \bar{y}_{i..} - \bar{y}_{i..} - \bar{y}_{i..} + 2\bar{y}_{...})^2$$

The product terms vanish, since the algebraic sum of deviations from mean is zero.
or, T.S.S. = S.S.R. + S.S.C. + S.S.T. + S.S.E.
where T.S.S. is the total sum of squares and S.S.R., S.S.C., S.S.T. and S.S.E. represent sum of squares due to rows, columns, treatments and error respectively, given by

$$T.S.S. = \sum_{i,j,k\in S} (y_{ijk} - \bar{y}_{...})^2 ; \quad S.S.R. = m\sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 ; \quad S.S.C. = m\sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$S.S.T. = m.\sum_k (\bar{y}_{..k} - \bar{y}_{...})^2 ; \quad S.S.E. = \sum_{i,j,k\in S} (y_{ijk} - \bar{y}_{i..} - \bar{y}_{i..} - \bar{y}_{i..} + 2\bar{y}_{...})^2$$

$$S.S.R = S_R^2 = m\sum_i \bar{Y}_{i..} - \bar{Y}...)^2 \qquad S.S.C = S_C^2 = m\sum_j Y_{.j.} - \bar{Y}...)^2$$

$$S.S.T = S_T^2 = m\sum_k \bar{Y}.k - \bar{Y}...)^2$$

S.S.E. = S^2_E = T.S.S. — S.S.R. — S.S.C. — S.S.T.

Sources of variations	Sum of squares	d.f	Mean sum of squares	Variance Ratio
Rows Columns Treatments Error	S.S.R $= S_R^2$ S.S.R $= S_R^2$ S.S.T $= S_T^2$ $E.S.S = S_E^2$	m - 1 m - 1 m - 1 $(m-1)(m-2)$	$s_R^2 = S_R^2/(m-1)$ $s_C^2 = S_C^2/(m-1)$ $s_T^2 = S_T^2/(m-1)$ $s_E^2 = S_E^2/(m-1)(m-2)$	$F_R = s_R^2/s_E^2$ $F_C = s_C^2/s_E^2$ $F_T = s_T^2/s_E^2$
Total	$T.S.S$	$m^2 - 1$		

Let us set up **Null Hypotheses**:

For row effects, $H_{o1}: \alpha_i = 0, \forall\, i$

For column effects, $H_{o2}: \beta_j = 0, \forall\, j$

For treatment effects, $H_{o3}: \gamma_k = 0, \forall\, k$

Alternative Hypotheses:

For row effects, $H_{11}: atleast\ one\ \alpha_i \neq 0$

For column effects, $H_{12}: atleast\ one\ \beta_j \neq 0$

For treatment effects, $H_{13}: atleast\ one\ \gamma_k \neq 0$

The variance ratios F_R, Fc &F_T follow (central) F distribution with
[(m - 1), (m - 1) (m - 2)] d.f. under the null hypotheses H_α H_β and H_τ respectively.

Let $F_a = F_a$ {(m - 1), (m - 1) (m - 2)) be the tabulated value of F for [(m - 1), (m - 1) (m - 2)] d.f. at the level of significance α. Thus if $F_R > F_\alpha$, we reject H_α and if $F_R \leq F_\alpha$, we may accept H_α. Similarly we can test and H_{02} and H_{03}.

Efficiency of LSD over RBD

Case-I When rows of LSD are taken as blocks, suppose that S^2_E be the error mean square for RBD with rows(of LSD) as blocks then efficiency of LSD relative to RBD is given by

$$E_1 = \frac{S^2{}_E'}{S^2{}_E} = \frac{s_c^2 + (m-1)s_E^2}{ms_E^2}$$

Case-II When columns of LSD are taken as blocks, suppose that $S^2{}_E$ be the error mean square for RBD with columns (of LSD) as blocks then efficiency of LSD relative to RBD is given by

$$E_2 = \frac{S^2{}_E''}{S^2{}_E} = \frac{s_R^2 + (m-1)s_E^2}{ms_E^2}$$

Efficiency of LSD relative to RBD : It is given by

$$E_3 = \frac{s_R^2 + s_C^2 + (m-1)s_E^2}{(m+1)s_E^2}$$

13.5 ESTIMATION OF MISSING VALUES IN LATIN SQUARE DESIGN.

Let us suppose that in Latin Square, the observation occurring in the ith row, jth column and receiving the kth treatment is missing. Let us assume that its value is x, i.e.,

R= Total of the known observations in the ith row, i.e., the row containing *'x'*.

C = Total of known observations in the jth column, i.e., the column containing *'x'*.

T = Total of known observations receiving kth treatment, i.e., total of all known treatment values containing '*x*'.

S= Total of all known observations. Then

$$SSR = \frac{(R+x)^2}{m} + Constant\ w.r.t.'\ x' - \frac{(S+x)^2}{m^2}$$

$$SSC = \frac{(C+x)^2}{m} + Constant\ w.r.t.'\ x' - \frac{(S+x)^2}{m^2}$$

$$SST = \frac{(T+x)^2}{m} + Constant\ w.r.t.'\ x' - \frac{(S+x)^2}{m^2}$$

$$TSS = x^2 + Constant\ w.r.t.'\ x' - \frac{(S+x)^2}{m^2}$$

$$E - SSE = TSS - SSR - SSC - SST$$

$$= x^2 - \frac{1}{m}[(R+x)^2 + (C+x)^2 + (T+x)^2] + 2\frac{(S+x)^2}{m^2}$$

We choose x so as minimize E

Therefore,

$$\frac{\partial E}{\partial x} = 0 \Rightarrow 2x - \frac{2}{m}(R + C + T + 3x) + \frac{4(S + x)}{m^2}$$

$$\Rightarrow \hat{x} = \frac{m(R + C + T) - 2S}{(m-1)(m-2)}$$

After inserting the estimated value for missing observation, we perform the usual analysis of variance, subtracting one d.f. for total S.S. and consequently for Error S.S. Adjusted treatment S.S. is obtained by subtracting the quantity

$\frac{[(m-1)T+R+C-S]^2}{[(m-1)(m-2)]^2}$ from the treatment S.S.

13.6 ILLUSTRATIONS

Example: An experiment was carried out to determine the effect of claying the ground on the field of barley grains; amount of clay used were as follows:

A : No clay

B: Clay at 100 per acre

C: Clay at 200 per acre

D : Clay at 300 per acre.

The yields were in plots of 8 metres by 8 metres and are given in Table

Column → / *Row* ↓	*I*	*II*	*III*	*IV*	*Row totals* (R_i)
I	*D* 29·1	*B* 18·9	*C* 29·4	*A* 5·7	83·1
II	*C* 16·4	*A* 10·2	*D* 21·2	*B* 19·1	66·9
III	*A* 5·4	*D* 38·8	*B* 24·0	*C* 37·0	105·2
IV	*B* 24·9	*C* 41·7	*A* 9·5	*D* 28·9	105·0
Column Totals (C_j)	75·8	109·6	84·1	90·7	360·2

(a) Perform the ANOVA (b) Calculate the efficiency of the above Latin Square Design over (i) R.B.D. and (ii) C.R.D. (c) Yield under 'A' in the first column was missing. Estimate the missing value and carry out the ANOVA.

Sol. The four treatment totals are :A = 30•8, B = 86•9, C= 1245, D = 1180 Grand total G = 3602, N = 16.

C.F. $=(360.2)^2/16 = 8109.025$

Raw S.S. $= (29.1)^2 + (18.9)^2 + ... + (9.5)^2 = (28.9)^2 = 10{,}052.08$

Total S.S = 10,052.08 - 8,109.0025 = 1,943.0775

$$\text{S.S.R.} = \frac{1}{4}\left[(83{\cdot}1)^2 + (66{\cdot}9)^2 + (105{\cdot}2)^2 + (105{\cdot}0)^2\right] - 8{,}109{\cdot}0025$$

$$= \frac{33{,}473{\cdot}26}{4} - 8{,}109{\cdot}0025 = 259{\cdot}3125$$

$$\text{S.S.C.} = \frac{1}{4}\left[(75{\cdot}8)^2 + (109{\cdot}6)^2 + (84{\cdot}1)^2 + (90{\cdot}7)^2\right] - 8{,}109{\cdot}0025$$

$$= \frac{33057{\cdot}10}{4} - 8109{\cdot}0025 = 155{\cdot}2725$$

$$\text{S.S.T.} = \frac{1}{4}\left[(30{\cdot}8)^2 + (86{\cdot}9)^2 + (124{\cdot}5)^2 + (118{\cdot}0)^2\right] - 8{,}109{\cdot}0025$$

$$= \frac{37924{\cdot}50}{4} - 8109{\cdot}0025 = 1372{\cdot}1225$$

$$\text{Error S.S.} = \text{T.S.S.} - \text{S.S.R.} - \text{S.S.C.} - \text{S.S.T.} = 156{\cdot}3700$$

Tabulated $F_{3,\,6}(0.05) = 4.76$

Source of variation	d.f.	S.S.	M.S.S.	Variance Ratio
(1)	(2)	(3)	(4) = (3) ÷ (2)	
Rows	3	259·5375	86·4375	$F_R = \frac{86·4375}{26·0616} = 3·32 < 4·76$
Columns	3	155·2725	51·7575	$F_c = \frac{51·7575}{26·0616} = 1·98 < 4·76$
Treatments	3	1,372·1225	457·3742	$F_T = \frac{457·3742}{26·0616} = 17·55 > 4·76$
Error	6	156·3700	26·0616	
Total	15	1,943·0775		

Hence we conclude that the variation due to rows and columns is not significant but the treatments, i.e., different levels of clay, have significant effect on the yield.

Efficiency of L.S.D.:(i) Relative efficiency of L.S.D. over R.B.D. when rows are taken as blocks is:

$$\frac{51·7575 + 3 \times 26·0616}{4 \times 26·0616} = 1·2465$$

Relative efficiency of L.S.D. over R.B.D. when columns are taken as blocks is

$$\frac{s_R^2 + (m-1)\, s_E^2}{m\, s_E^2} = \frac{86·4375 + 3 \times 26·0616}{4 \times 26·0616} = 1·5792$$

(iii) Relative efficiency of an L.S.D. over C.R.D. is:

$$\frac{s_R^2 + s_C^2 + (m-1)s_E^2}{(m+1)\, s_E^2} = \frac{116·3798}{130·3080} = 1·6605$$

Hence, the gain by using Latin Square Design

(i) instead of R.B.D. is 25% when rows are taken as blocks and 58% when columns are taken as blocks.

(ii) instead of C.R.D. is 66%.

13.7 SUMMARY

*** Latin Square Design**: In this design the number of treatments is equal to the number of replications. Thus in case of m treatments, there have to be $m \times m = m^2$ experimentals

units. The whole of the experimental material is divided into m^2 experimental units arranged in a square so that each row as well as each column contains m units. The m treatments are then allocated at random to these rows and columns in such a way that every treatment occurs once and only once in each row and in each column. Such a layout is known as $m \times m$ latin square design.

* **Model of LSD**: If a single observation is made per experimental unit then the linear model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijk},$$

where μ is the constant mean effect; α_i, β_j *and* γ_k are the constant effects due to the ith row, jth column and kth treatment respectively and e_{ijk} is error effect due to random component assumed to be normally distributed with mean zero and variance σ_e^2 i.e., $e_{ijk} \overset{i.i.d}{\sim} N(0, \sigma_e^2)$.

* **Missing value in LSD:** Let x be the missing value occurring in the ith row, jth column and receving kth treatment, then

$$\hat{x} = \frac{m(R + C + T) - 2S}{(m-1)(m-2)}$$

where,

R= Total of the known observations in the ith row, i.e., the row containing '*x*'.

C = Total of known observations in the jth column, i.e., the column containing '*x*'.

T = Total of known observations receiving kth treatment, i.e., total of all known treatment values containing '*x*'.

S = Total of all known observations.

***Relative efficiency of LSD over RBD**

i) when rows are taken as blocks is $\frac{s_C^2 + (m-1)s_E^2}{ms_E^2}$

ii) when columns are taken as blocks $\frac{s_R^2+(m-1)s_E^2}{ms_E^2}$

and relative efficiency of LSD over CRD is $\frac{s_R^2+s_C^2+(m-1)s_E^2}{(m+1)s_E^2}$

FACTORIAL EXPERIMENT

STRUCTURE

14.1 Factorial Experiment

14.2 Advantages of Factorial Experiment

14.3 2^2 Factorial Design

14.4 Statistical Analysis of 2^2 Factorial Design

14.5 2^3 Factorial Design

14.6 Statistical Analysis of 2^3 Factorial Design

14.7 2^n Factorial Design

14.8 Analysis of 2^n Factorial Design

14.9 Yates method of computing Factorial Effect Total

14.1 FACTORIAL EXPERIMENTS:

We may have to face the situation e.g., In industrial applications where quite often several factors may affect the characteristics in which we are interested, and we wish to estimate the effects of each of the factors and how the effect of one factor varies over the level of the other factors. For example, the yield of a chemical process may be affected by several factors such as the levels of pressure, temperature, irrigation etc. One might try to test each of the factors separately, holding all other factors constant in a given experiment, but from such an experiment might not give the information required. The logical procedure

would be to vary all factors simultaneously, i.e., within the framework of the same experiment. When we do so, we have what is now widely known as a factorial experiment.

In C.R.D. or R.B.D. or L.S.D., we were primarily concerned with the comparison and estimation of the effects of a single set of treatments like varieties of wheat, manure of different methods of cultivation etc. Such experiments which deal with one factor only may be called simple experiments. In factorial experiment as the adjective factorial indicates, the effects of several factors of variation are studied and investigated simultaneously, the treatments being all the combinations of different factors under study. In these experiments an attempt is made to estimate the effects of each of the factors and also the interaction effects, i.e., the variation in the effect of one factor as a result to different levels of other factors.

As a simple illustration let us consider two fertilizers, say, Potash (K) and Nitrogen (N). Let us suppose that there are p different varieties of Potash and q different varieties of Nitrogen. p and q are termed as the levels of the factors potash and Nitrogen respectively. To test the effectiveness of various treatments, viz., different levels of Potash or Nitrogen we have to conduct two simple experiments, one for Potash and the other for Nitrogen. A series of experiments in which only one factor is varied at a time would be both lengthy and costly and might still be unsatisfactory. Moreover, these simple experiments do not give us any information interaction effect.

The only alternative is to try to investigate the variations in several factors simultaneously by conducting the above experiment as a p x q factorial experiment, where p and q are the levels of various factors under consideration.

In general, if levels of various factors are equal then s^n factorial experiment means an experiments with the n factors, each at s levels where n is any positive integer greater than or equal to 2, e.g., 2^3 means 3 factors, each at 2 levels

14.2 ADVANTAGES OF FACTORIAL EXPERIMENT.

1. It increases the scope of the experiment and its inductive value and also giving information not only on the main factors but on their interactions.

2. The various levels of one factor constitute replications of other factors and increase the amount of information obtained on all factors.

3. When there are no interactions, the factorial design gives the maximum efficiency in the estimate of the effects.

4. When interactions exist, their nature being unknown a factorial design is necessary to avoid misleading conclusions.

5. In the factorial design the effect of a factor is estimated at several levels of other factors and the conclusions hold over a wide range of conditions.

Contrast: A linear combination $\sum_{i=1}^{k} c_i t_i$ of k treatment means t_i (i = 1, 2, ..., k) is a contrast (or a comparison) of treatment means if $\sum_{i=1}^{k} c_i = 0$. In other words, contrast is a linear combination of treatment means such that the sum of the coefficients is zero.

Orthogonal Contrasts: Two contrasts of k-treatment means t_i (i = 1, 2, ...,k), viz., are said to be orthogonal if

$\sum_{i=1}^{k} c_i d_i = 0$, $\sum_{i=1}^{k} c_i t_i$, $\sum_{i=1}^{k} c_i = 0$ and $\sum_{i=1}^{k} d_i t_i$, $\sum_{i=1}^{k} d_i = 0$ **are said to be**

In other words, the contrasts are orthogonal if the sum of the product of the coefficients of corresponding treatment means is zero.

14.3 2^2-FACTORIAL DESIGN.

Here we have two factors each at two levels (0,1), say, so that there are $2 \times 2 = 4$ treatment combinations in all. Following the notations due to Yates, let the capital letters A and B indicate the names of the two factors under study and let the small letters a and b denote one of the two levels of each of the corresponding factors and this will be called the second level. The first level of A and B is generally expressed by the absence of the corresponding letter in the treatment combinations. The four treatment combinations can be enumerated as follows

a_0b_0 or '1' : Factors A and B, both at first level.

a_1b_0 or a : A at second level and B at first level.

a_0b_1 or b : A at first level and B at second level.

a_1b_1 or ab : A and B both at second level.

These four treatment combinations can be compared by laying out the experiment in R.B.D., with r replicates (say), each replicate containing 4 units, or 4 x 4 L.S.D., and ANOVA can be carried out accordingly. In the above cases there are 3 d.f. associated with the treatment effects. In factorial experiment our main objective is to carry out separate tests for the main effects A, B and the interaction AB, splitting the treatment S.S. with 3 d.f. into three orthogonal components each with 1 d.f. and each associated either with the main effects A and B or the intersection AB.

Main Effects and Interactions. Suppose the factorial experiment with $2^2 = 4$ treatments is conducted in r-blocks or replicates as they are often called. Let [1], [a], [bl and [ab] denote the total yields of the r-units (plots) receiving the treatments 1, a, b and ab respectively and let the corresponding mean values obtained on dividing these totals by r be denoted by (1), (a), (b) and (ab) respectively.

The effect of factor A can be represented by the difference between mean yields obtained at each level.

Thus the effect of factor A at the first level b_0 of B = $(a_1b_o) - (a_ob_o) = (a) - (1)$……(1)

the effect ofA at the second level b_1 of B = $(a_1b_1) - (aob_1) = (ab) - (b)$ ……(2)

The average observed effect of A over the two levels of B is called the main effect due to A and is defined by

$$A = \frac{1}{2}[(ab) - (b) + (a) - (1)] = \frac{1}{2}[(a-1)(b+1)]$$

Similarly we shall get the main effect due to factor B as

$$B = \frac{1}{2}[(ab) + (b) - (a) - (1)] = \frac{1}{2}[(a+1)(b-1)]$$

The interaction of two factors is the failure of the levels of one factor, say, A to retain the same order and magnitude of performance throughout all levels of the second factor, say, B. If the two factors act independently of one another, we should expect the true effect of

one to the same at either level of other.

In other words, we should expect that the two expressions (1) and (2) were really the estimates of the same thing .On the other hand if two factors are not independent, as a measure of the extent to which the factors interact and we write the two-factor interaction or the first order interaction between the factors A and B as

$$AB = \frac{1}{2}[(ab) - (b) - (a) + (1)] = \frac{1}{2}[(a-1)(b-1)]$$

14.4 STATISTICAL ANALYSIS OF 2^2-DESIGN.

Factorial experiments are conducted either in C.R.D. R.B.D. or L.S.D. and thus they can be analysed in the usual manner except that in this ease the treatment S.S. is split into three orthogonal components each with 1 d.f. .The S.S. due to the factorial effects A, B and AB is obtained by multiplying the squares of the factorial effects by a suitable quantity. In practice, these effects are usually computed from the treatment totals [a], [b], [ab] etc., rather than from the treatment means (a), (b), etc. Factorial effect totals are given by the expressions

$$[A] = [ab] - [b] + [a] - [1]$$

$$[B] = [ab] + [b] - [a] - [1]$$

$$[AB] = [ab] - [b] - [a] + [1]$$

S.S. due to any factorial effect is obtained on multiplying the square of the effect total by factor (1/4r), where r is the common replication number Thus

S.S. due to main effect A=$[A]^2/4r$

S.S. due to main effect B=$[B]^2/4r$

S.S. due to interaction AB=$[AB]^2/4r$

each with 1 d.f

ANOVA TABLE FOR FIXED EFFECT MODEL TWO FACTOR (2^2) EXPERIMENT IN R.B.D. IN 'r' REPLICATES

Sources of variation	d.f	S.S.	M.S.S.	Variance ratio 'F'
Blocks or replicates	r-1	SSB	$s_b^2 = \frac{SSB}{r-1}$	$F_b = \frac{s_b^2}{s_E^2}$
Treatment	3	SST	$s_T^2 = \frac{SST}{3}$	$F_T = \frac{s_T^2}{s_E^2}$
Main Effect A	1	S_A^2	$s_A^2 = \frac{S_A^2}{1}$	$F_A = \frac{s_A^2}{s_E^2}$
Main Effect B	1	S_B^2	$s_B^2 = \frac{S_B^2}{1}$	$F_B = \frac{s_B^2}{s_E^2}$
Interaction Effect AB	1	S_{AB}^2	$s_{AB}^2 = \frac{S_{AB}^2}{1}$	$F_{AB} = \frac{s_{AB}^2}{s_E^2}$
Error	3(r-1)	SSE	$s_E^2 = \frac{SSE}{3(r-1)}$	
Total	4r-1	TSS		

Here each of the statistics F_A, F_B and F_{AB} follows central F-distribution with [1, 3 (r-1)] d.f If for any factorial effect, calculated F is greater than tabulated F for [1, 3(r-1)] d.f. and at certain level of significance 'say' α , then the null hypothesis H_0 of the presence of the factorial effect is rejected, other-wise H_0 may be accepted**.**

14.5 2^3-FACTORIAL EXPERIMENT.

In 2^3-experiment we consider three factors, say, A, B, and C each at two levels, say, (a_o, a_1), (b_0, b_1) and (c_o, c_1) respectively, so that there are $2^3 = 8$ treatment combinations in all. Let the corresponding small letters a, b and c denote the second level of each of the corresponding factors. The first level of each factor A, B and C is signified by the absence of the corresponding letter in the treatment combinations. The eight treatment combinations in a standard order are

'1', a, b, ab, c, ac, bc, abc,

This factorial experiment can be performed as a C.R.D. with 8 treatments, or R.B.D. with r replicates (say), each replicate containing 8 treatments of L.S.D. with m = 8 and data can be analysed accordingly. Here we split up the treatment S.S. with 7 d.f. into 7 orthogonal components corresponding to the three main effects A, B and C, three first order (or two factor) interactions AB, AC, and BC and one second order interaction (or three factor interaction) ABC, each carrying 1 d.f.

Main Effects and Interactions. The simple effect of A, (say), is given by the differences in the mean yields of A as a result of increasing the factor A from the level a_0 to a_1, at other levels of the factors B and C.

Level of B	*Level of C*	*Simple effect of A*
b_0	c_0	$(a_1b_0c_0) - (a_0b_0c_0) = (a) - (1)$
b_1	c_0	$(a_1b_1c_0) - (a_0b_1c_0) = (ab) - (b)$
b_0	c_1	$(a_1b_0c_1) - (a_0b_0c_1) = (ac) - (c)$
b_1	c_1	$(a_1b_1c_1) - (a_0b_1c_1) = (abc) - (bc)$

.....(1)

The main effect of A is defined as the average of these 4 simple effects. Thus

$$A = \frac{1}{4}[(abc) - (bc) + (ac) - (c) + (ab) - (b) + (a) - (1)]$$

$$= \frac{1}{4}[(a-1)(b+1)(c+1)]$$

Similarly the main effects of the factors B and C can be obtained to give

$$B = \frac{1}{4}[(abc) + (bc) - (ac) - (c) + (ab) + (b) - (a) - (1)]$$
$$= \frac{1}{4}[(a+1)(b-1)(c+1)]$$

$$C = \frac{1}{4}[(abc) + (bc) + (ac) + (c) - (ab) - (b) - (a) - (1)]$$
$$= \frac{1}{4}[(a+1)(b+1)(c-1)]$$

If the factors A and B were independent then we would expect that the average effect of A will remain the same at either level of B. In this case mean of these two average effects will give the main effect of A, as obtained in (1).But if the factors A and B are not independent,

then a measure of their interaction AB is given by half of the difference between the average effect of A at the second and the first level of B. Symbolically, we have

$$AB = \frac{1}{4}[(abc) - (bc) - (ac) - (c) + (ab) - (b) + (a) - (1)]$$
$$= \frac{1}{4}[(a-1)(b-1)(c+1)]$$

Similarly we can obtain expressions for the interactions BC and AC as given below

$$BC = \frac{1}{4}[(a-1)(b-1)(c+1)]$$

$$AC = \frac{1}{4}[(a-1)(b+1)(c-1)]$$

Similarly we can obtain expressions for the second order interaction as given below

$$ABC = \frac{1}{4}[(abc) - (bc) - (ac) + (c) - (ab) + (b) + (a) - (1)]$$
$$= \frac{1}{4}[(a-1)(b-1)(c-1)]$$

14.6 STATISTICAL ANALYSIS OF 2^3-DESIGN.

By using the table of divisors and signs of a 2^3- factorial experiment or Yates' method the various factorial effect totals can be expressed as mutually orthogonal contrasts of the 8 treatment totals. Thus, e.g.,

[A]=[abc]-[bc]+[ac]-[c]+[ab]-[b]+[a]-[a]

and so on. In the analysis of 2^3-design we split the treatment S.S. with 7 d.f. into 7 mutually orthogonal components corresponding to seven factorial effects, each carrying 1 d.f.

The S.S. due to any of the factorial effect is given by

$$\frac{[\cdot]^2}{8r}$$

i.e., S.S. due to any factorial effect, main or interaction is obtained on multiplying the

square of the factorial effect total by 1/(8r), where r is the common replication number. Thus, for example

S.S. due to any main effect A=$\frac{[A]^2}{8r}$ with 1 d.f.

and so on

ANOVA can now be carried out as below given in the tabular form

Source of Variation	*d.f.*	*S.S.*	*M.S.S.* = $\frac{S.S.}{d.f.}$	*Variance Ratio 'F'*
Replications (Blocks)	$(r-1)$	S_R^2	$s_R^2 = \frac{S_R^2}{r-1}$	$F_R = \frac{S_R^2}{S_E^2} \sim F[r-1, 7(r-1)]$
Main Effects				
A	1	$S_A^2 = [A]^2/8r$	$s_A^2 = S_A^2$	$F_A = s_A^2/s_E^2 \sim F[1, 7(r-1)]$
B	1	$S_B^2 = [B]^2/8r$	$s_B^2 = S_B^2$	$F_B = s_B^2/s_E^2 \sim F[1, 7(r-1)]$
C	1	$S_C^2 = [C]^2/8r$	$s_C^2 = S_C^2$	$F_C = s_C^2/s_E^2 \sim F[1,7(r-1)]$
1st Order Interactions				
AB	1	$S_{AB}^2 = [AC]^2/8r$	$s_{AB}^2 = S_{AB}^2$	$F_{AB} = s_{AB}^2/s_E^2 \sim F[1,7(r-1)]$
AC	1	$S_{AC}^2 = [AC]^2/8r$	$s_{AC}^2 = S_{AC}^2$	$F_{AC} = s_{AC}^2/s_E^2 \sim F[1, 7(r-1)]$
BC	1	$S_{BC}^2 = [BC]^2/8r$	$s_{BC}^2 = S_{BC}^2$	$F_{BC} = s_{BC}^2/s_E^2 \sim F[1,7(r-1)]$
2nd Order Interaction				
ABC	1	$S_{ABC}^2 = [ABC]^2/8r$	$s_{ABC}^2 = S_{ABC}^2$	$F_{ABC} = s_{ABC}^2/s_E^2 \sim F[1, 7(r-1)]$
Error	$7(r-1)$	S_E^2 = By subtraction	$s_E^2 = \frac{S_E^2}{7(r-1)}$	
Total	$r.2^3 - 1 = 8r - 1$			

The hypothesis of the presence of a factorial effect is rejected at α % level of significance if the corresponding calculated F-statistic in the table is greater than tabulated $F_{1,7(r-1)}$ otherwise the hypothesis may be accepted.

14.7 2^N-FACTORIAL EXPERIMENT.

In order to analyse 2^n factorial we can generalized the notations and results of 2^2 and 2^3 factorial experiment.

Here we consider n factors each at 2 levels. Suppose A, B, C, D, ..., K are the factors each at two levels (0, 1). Corresponding small letters a, b, c, d, ..., k denote the corresponding factors at the second level, the first level of any factor being signified by the absence of the corresponding small letter. The treatment combinations, in standard order, can be written as

1, a, b, ab, c, ac, be, abc, d, ad, bd, abd, ed, acd, bed, abcd,e, ae, be, abe, ce, ace, bce, abee, de, ade, bde, abde, ede, acde, bede, abcde, etc.

For 2^n-experiment, the various factorial effects are enumerated as follows

Main effects : $\binom{n}{1}$ in number

Two factor interactions: $\binom{n}{2}$ in number

Three factor interactions: in number

n factor interactions: in number

Hence, the total number of factorial effects in 2^n-experiment are

$$^nC_1 + {}^nC_2 + \ldots + {}^nC_n = [{}^nC_0 + {}^nC_1 + \ldots + {}^nC_n] - 1 = (1+1)^n - 1 = 2^n - 1.$$

Main Effects and Interactions. As in the case of 2^2 and 2^3-experiments the results for main effects and interactions can be generalised to the case 2^n-experiment. Thus, for n factors A, B, C, D, ..., K, the main effects and interactions are given by the expression:

$$\frac{1}{2^{n-1}}[(a \pm 1)(b \pm 1)(c \pm 1) \ldots (k \pm 1)]$$

the corresponding sign in each factor being taken as negative if the corresponding factor is contained in the factorial effect whose value we want. The factorial effect totals can be obtained every conveniently from treatment totals by the generalisation of F. Yates' method for 2^2 and 2^3 experiments. As pointed out there, for 2^n experiment we shall need n cycles of the 'sum and difference' procedure.'

14.8 ANALYSIS OF 2^N DESIGN.

It will be seen that all the factorial effects (main and interaction) are mutually orthogonal contrasts of treatment totals. Hence, having obtained the factorial effect totals by Yates' technique, the S.S. due to each factorial effect is given by:

$$\frac{[\cdot]^2}{2^n r}$$

where [·] is the factorial effect total.

ANOVA TABLE FOR FACTORIAL EXPERIMENT IN R RANDOMISED BLOCKS

Source of Variation	*d.f.*	*S.S.*	*M.S.S*
Blocks	$r-1$	$S_R^2 = \frac{\sum B_j^2}{2^n} - \text{C.F.}$	$s_R^2 = \frac{S_R^2}{r-1}$
Treatments	$2^n - 1$	$S_T^2 = \frac{\sum T_i^2}{r} - \text{C.F.}$	$s_T^2 = \frac{s_T^2}{2^n - 1}$
Main effects			
A	1	$S_A^2 = [A^2/r.\ 2^n$	$s_A^2 = S_A^2$
B	1	$S_B^2 = [B]^2/r.2^n$	$s_B^2 = S_B^2$
⋮	⋮	⋮	⋮
K	1	$S_K^2 = [K]^2 / r.2^n$	$s_K^2 = S_K^2$
Two-factor Interactions			
AB	1	$S_{AB}^2 = [AB]^2 / r.\ 2^n$	$s_{AB}^2 = S_{AB}^2$
AC	1	$S_{AC}^2 = [AC]^2 / r.\ 2^n$	$s_{AC}^2 = S_{AC}^2$
BC	1	$S_{BC}^2 = [BC]^2 / r.2^n$	$s_{BC}^2 = S_{BC}^2$
⋮	⋮	⋮	⋮
Three-factor Interactions			
ABC	1	$S_{ABC}^2 = [ABC]^2 / r.2^n$	$s_{ABC}^2 = S_{ABC}^2$
ACD	1	$S_{ACD}^2 = [ACD]^2 / r.2^n$	$s_{ACD}^2 = S_{ACD}^2$
⋮	⋮	⋮	⋮
n -factor interaction $ABCD...K$	1	$S_{AB...K}^2 = [AB...K]^2 / r.2^n$	$s_{AB...K}^2 = S_{AB...K}^2$
Error	$(r-1)(2^n-1)$	S_E^2 = By subtraction	$s_E^2 = \frac{S_E^2}{(r-1)(2^n-1)}$
Total	$r.\ 2^n - 1$	Raw S.S. – C.F.	

14.9 YATES' METHOD OF COMPUTING FACTORIAL EFFECT TOTALS.

For the calculation of various factorial effect totals for 2^n factorial experiments F. Yates developed a special computational rule which enables us to avoid specific algebraic formulae. Yates' method consists in the following steps :

1. In the first column we write the treatment combinations. It is an essential part of the procedure that the treatment combinations be written in a standard systematic order as explained below:

"Starting with the treatment combination 1, each factor is introduced in turn and is then followed by all combinations of itself with the treatment combinations previously written down, e.g., for 2^2-experiment with factors A and B, the order of treatment combinations will be 1, a, b, ab and and so on

2. Against each treatment combination, write the corresponding total yields from all the replicates.
3. The entries in the third column can be split into two halves. The first half is obtained by writing down in order, the pairwise sums of the values in column 2 and the second half is obtained by writing in the same order the pairwise differences of the values in second column.
4. To complete the next (4th) column, the whole of the procedure as explained in step 3 is repeated on column 3, and for 2^3-design, the 5th column is derived from 4th in a similar manner.

Yates Method for 2^2- Experiment

Treatment Combination (1)	*Total Yield from all replicates (2)*	*(3)*	*(4)*	*Effect Totals*
'1'	[1]	[1] + [a]	[1] + [a] + [b] + [ab]	Grand Total
a	[a]	[b] + [ab]	[ab] – [b] + [a] – [1]	[A]
b	[b]	[a] – [1]	[ab] + [b] – [a] – [1]	[B]
ab	[ab]	[ab] – [b]	[ab] – [b] – [a] + [1]	[AB]

Thus for a 2^n-factorial experiment there will be *n* cycles of this "***sum and difference***" procedure. The first term in the last, viz., (n + 2)th column always given the grand total (G) while the other entries in the last column are the totals of the main effects or the interactions corresponding to the treatment combinations in the first column of the table. we illustrate Yates' Method for 2^2 and 2^3 factorial experiments respectively.

Yates Method for 2^3- Experiment

Treatment Combination (1)	*Treatment Totals* (2)	*(3)*	*(4)*	*(5)*	*Effect Totals*
'1'	[1]	$[1] + [a] = u_1$ (say)	$u_1 + u_2 = v_1$	$v_1 + v_2 = w_1$	Grand Total
a	$[a]$	$[b] + [ab] = u_2$ (say)	$u_3 + u_4 = v_2$	$v_3 + v_4 = w_2$	$[A]$
b	$[b]$	$[c] + [ac] = u_3$ (say)	$u_5 + u_6 = v_3$	$v_5 + v_6 = w_3$	$[B]$
ab	$[ab]$	$[bc] + [abc] = u_4$ (say)	$u_7 + u_8 = v_4$	$v_7 + v_8 = w_4$	$[AB]$
c	$[c]$	$[a] - [1] = u_5$ (say)	$u_2 - u_1 = v_5$	$v_2 - v_1 = w_5$	$[C]$
ac	$[ac]$	$[ab] - [b] = u_6$ (say)	$u_4 - u_3 = v_6$	$v_4 - v_3 = w_6$	$[AC]$
bc	$[bc]$	$[ac] - [c] = u_7$ (say)	$u_6 - u_5 - v_7$	$v_6 - v_5 = w_7$	$[BC]$
abc	$[abc]$	$[abc] - [bc] = u_8$ (say)	$u_8 - u_7 = v_8$	$v_8 - v_7 = w_8$	$[ABC]$