Indira Gandhi National Open University
School of Sciences

MST-005
Statistical Techniques

Block

1

SAMPLING DESIGNS

UNIT 1
Introduction to Sample Surveys **7**

UNIT 2
Simple Random Sampling **25**

UNIT 3
Stratified Random Sampling **45**

UNIT 4
Some Other Sampling Schemes **61**

Curriculum and Course Design Committee

Prof. K. R. Srivathsan
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Parvin Sinclair
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Geeta Kaicker
Director, School of Sciences
IGNOU, New Delhi

Prof. R. M. Pandey
Department of Bio-Statistics
All India Institute of Medical Sciences
New Delhi

Prof. Jagdish Prasad
Department of Statistics
University of Rajasthan, Jaipur

Prof. Rahul Roy
Mathematics and Statistics Unit
Indian Statistical Institute, New Delhi

Dr. Diwakar Shukla
Department of Mathematics and Statistics
Dr. Hari Singh Gaur University, Sagar

Prof. G. N. Singh
Department of Applied Mathematics
I. S. M. University, Dhanbad

Prof. Rakesh Srivastava
Department of Statistics
M. S. University of Baroda, Vadodara

Dr. Gulshan Lal Taneja
Department of Mathematics
M. D. University, Rohtak

Faculty Members, School of Sciences, IGNOU

Statistics
Dr. Neha Garg
Dr. Nitin Gupta
Mr. Rajesh Kaliraman
Dr. Manish Trivedi

Mathematics
Dr. Deepika Garg
Prof. Poornima Mital
Prof. Sujatha Varma
Dr. S. Venkataraman

Block Preparation Team

Content Editor
Prof. Jagdish Prasad
Department of Statistics
University of Rajasthan, Jaipur

Language Editor
Dr. Parmod Kumar
School of Humanities, IGNOU

Secretarial Support
Mr. Deepak Singh

Course Writer
Dr. Manish Trivedi
School of Sciences, IGNOU

Formated By
Dr. Manish Trivedi
Mr. Prabhat Kumar sangal
School of Sciences, IGNOU

Programme and Course Coordinator: Dr. Manish Trivedi

Block Production

Mr. Y. N. Sharma, SO (P.)
School of Sciences, IGNOU

Acknowledgement: We gratefully acknowledge Prof. Geeta Kaicker, Director, School of Sciences for her great support and guidance.

December, 2011

ISBN – 978-81-266-5784-1

Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by Director, School of Sciences.

Laser Type set by: Rajshree Computers, V-166A, Bhagwati Vihar, (Near Sector-2, Dwarka), Uttam Nagar, New Delhi-110059

Printed at: Gita Offset Printers Pvt. Ltd., C-90, Okhla Industrial Area, Phase-I, New Delhi-110020.

STATISTICAL TECHNIQUES

In MST-004, you have studied some of the sampling distributions, methods of estimation and various kinds of parametric and non parametric tests. In Block 1 of this course, we have discussed some frequently used methods of sampling with their characteristics and applications.

In Block 3 of MST-004, we have restricted ourselves to the test of equality of means of two populations only. If the number of populations are more than two and some one is interested to test the hypothesis of equality of means then the ANOVA test serves the purpose. In Block 2, we have discussed the mythology and applications of One-way and Two-way analysis of variance.

In various field of experimentation, we have to plan an experiment and design the execution without loss of time and energy. In Block 3, we have elaborated different kind of designs with, their layout and the statistical analysis.

In case, where the experiment results are different in nature and not good enough to reach on any decision, the need of simulation technique arises. Generation of the random numbers is an important part of simulation technique. In Block 4, we have discussed various kind of systems and methods of generation of the random numbers for discrete and continuous variables. The simulation techniques and their applications in different fields are also discussed in this block.

Notations and Symbols

N	:	Population size / Number of units or elements in the population
n	:	Sample size / Number of units in sample
X_i	:	i^{th} unit or member in population
$\bar{X}$ or μ	:	Population mean
$\bar{x}$	:	Sample mean
S^2	:	Population mean square
s^2	:	Sample mean square
σ^2	:	Population variance
${}^{N}C_{n}$	:	Number of combinations of n units out of N units.
$E(\bar{x})$	:	Expected value of $\bar{x}$
$Var(\bar{x})$	:	Variance of $\bar{x}$
$Var(\bar{x}_{st})_{PROP}$	:	Variance of stratified sample mean under proportional allocation
$Var(\bar{x}_{st})_{NEY}$	:	Variance of stratified sample mean under Neyman's allocation
$Var(\bar{x}_{sys})$	:	Variance of systematic sample mean
S_{sys}^2	:	Population mean square of systematic sampling
SRSWR	:	Simple random sampling with replacement
SRSWOR	:	Simple random sampling without replacement
A	:	Number of population units possessing attribute A
A'	:	Number of population units not possessing attribute A
π	:	Population proportion
a	:	Number of sample units possessing attribute A
a'	:	Number of sample units not possessing attribute A
p	:	Sample proportion
SE (x)	:	Standard error of x
α	:	Level of significance
d	:	Difference between population mean and its estimate
t_α	:	Significant value of t at α level of significance

SAMPLING DESIGN

A sample survey has now become to be considered an organized fact finding instrument. Its importance to modern civilization lies in fact that it can be used to summarize the facts which would otherwise be in accessible owing to the remoteness and obscurity of the persons or to her unit concerned. Sample survey allows to make decisions to be made which take into account the significant factors of the problems they are meant to solve.

The information on a population may be collected in two ways. Either every unit in the population is enumerated which is called census or enumerated limited to only a part or a sample selected from the population called sample survey. A sample survey will usually be less costly and less time consuming than a complete census.

The main objective of this block is to present the theory and techniques of sample surveys with their applications. Sample surveys are to be widely used as a means of collecting information on to meet a definite need in government, industry and trade, physical and life sciences and technology, social, educational and economical problems, etc. All the walks of life are covered by sample surveys.

In Unit 1, a general introduction of the sample survey has been elaborated. In that unit, the basic principles, principle steps and types of sampling have been described. In Unit 2, we shall discuss the simple random sampling and its methodology. The properties of the simple random sampling are also described. In Unit 3, the stratified random sampling and its basic properties are discussed and in Unit 4 some other random sampling i.e. systematic random sampling, cluster sampling and two stage sampling with their basic properties are discussed.

Suggested Readings:

1. Goon, A. M., Gupta, M. K. and Das Gupta, B.; Fundamentals of Statistics, Vol II, World Press, Calcutta.
2. Gupta, S. C. and Kapoor, V. K.; Fundamentals of Applied Statistics, Sultan Chand & Sons.
3. Cochran, W. G.; Sampling Techniques (Chs. 13, 5-8, 10-13), John Wiley, 1963 and Wiley Eastern.
4. Deming, W. E.; Some Theory of Sampling (Chs. 1, 2, 4-6.), John Wiley, 1950.
5. Raj, D.; Sampling Theory, McGraw-Hill, 1968 and Tata McGraw-Hill.
6. Murthy, M. N.; Sampling Theory and Methods (Chs. 1-3, 5, 7, 9-11, 13-15), Statistical Publishing Society.
7. Sukhatme, P. V. and Sukhatme, B. V.; Sampling Theory of Surveys with Applications, FAO (United Nations) and Asia Publishing House, 1970.
8. Yates, F.; Sampling Methods in Censes and Surveys (Chs. 1-3, 6-8), Charles Griffin, 1960.

Notations and Symbols

N_i	:	Number of units in i^{th} stratum
n_i	:	Number of sample units selected from i^{th} stratum
X_{ij}	:	Value of the character under study for the j^{th} unit in i^{th} stratum
x_{ij}	:	Value of j^{th} sample unit taken from i^{th} stratum
$\overline{X}_i$	:	Mean of i^{th} stratum in population
$\overline{X}$	:	Population mean
W_i	:	Weight of i^{th} stratum
S_i^2	:	Population mean square of i^{th} stratum
$\overline{x}_i$	:	Sample mean of i^{th} stratum
s_i^2	:	Sample mean square of units selected from i^{th} stratum
$\overline{x}_{st}$	:	Stratified sample mean
π_i	:	Proportion of population units belonging to attribute A in i^{th} stratum
p_i	:	Proportion of sample units belonging to attribute A from i^{th} stratum
$Var(\overline{x}_{st})$	:	Variance of stratified sample mean
c_i	:	Cost per unit of i^{th} stratum
c_0	:	Over head fixed cost
C	:	Total cost
λ	:	Lagrange's multiplier
ρ	:	Intra-cluster correlation coefficient

UNIT 1 INTRODUCTION TO SAMPLE SURVEYS

Structure

1.1 Introduction
Objectives
1.2 Introduction to Population
1.3 Census
1.4 Sample Survey
1.5 Principles of Sample Survey
1.6 Principle Steps in Sample Survey
1.7 Sampling and Non-sampling Error
1.8 Advantages of Sampling over Census
1.9 Types of Sampling
1.10 Objectives of Sampling
1.11 Problems of Sampling Methods
1.12 Summary
1.13 Solutions / Answers

1.1 INTRODUCTION

The use of sampling in making inferences about a population is possible and has been in operation right from beginning. When one has to make an inference about a lot of large size and it is not practicable to examine each individual unit, then few units of the lot are examined and on the basis of the information of those units, one makes decisions about whole lot. For example, a person would like to purchase a bag of rice may examine a handful of rice from the bag and on the basis of that he/she makes his/her decision about the purchase of full bag.

A brief introduction to population, sample and sampling is given in Section 1.2. Census and sample survey are explored in Sections 1.3 and 1.4. The basic principles of sample survey are explained in Section 1.5 whereas the principle steps in sample survey are described in Section 1.6. The basic concepts of sampling and non-sampling error are provided in Section 1.7 and advantages of sampling over census are explained in Section 1.8. Various types of sampling methods are explored in Section 1.9 whereas the objectives of sampling are described in Section 1.10. Problems in sample surveys are dicussed in Section 1.11.

Objectives

After studying this unit, you would be able to

- define a population and explain the different kinds of population;
- describe the census and sample survey;
- describe the conditions and principles of sample survey;
- explain the principle steps in sample survey;

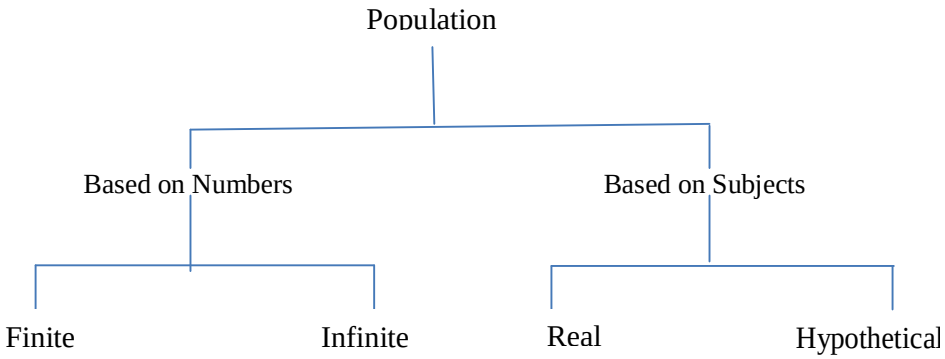
- describe the sampling and non-sampling error;
- explain the different kinds of sampling procedures; and
- explain the objectives, problems and importance of sampling.

1.2 INTRODUCTION TO POPULATION

A group of individuals having same characteristics in same surrounding is known as population. According to A. C. Rosander "A population is the totality of objects under consideration". In short, group of all objects which are coming under the definition of investigation unit. For example, the group of employees of a institute shall be the population for every investigation related to the employees. Individuals of the population are known as a unit or a element in population.

1.2.1 Types of Population

There are two types of a population as discussed below:



Finite or Infinite Population
In a finite population the number of units is finite i.e. number of students in a class, number of employees in a college. In a infinite population, the number of units are infinite, for example the number of hair in head etc. and also if the units of a population are unable to count and the limits cannot be made then the population is known as Infinite population, for example stars in sky, leaves in tree, number of viewers of a T.V. serial, etc.

Real or Hypothetical Population
Population of a concrete subject is called real population. For example, students in a university, employees of an institution. In a hypothetical population the subjects are not concrete, whether they are hypothetical. For example, the population made by the number of Heads or Tails based on the trial of a coin.

1.2.2 Information about Population

The need for adequate and reliable data is ever increasing for taking wise decisions in different fields of human activity and business. There are two ways in which the required information may be obtained:

1. Complete Enumeration or Census
2. Sample Survey

E 1) Define a population and describe the different types of populations.

1.3 CENSUS

In census, we study about each and every unit of the population. Population means total units of investigation area. In census, whole group related to investigation is investigated and the informations are collected, i.e. Census of population of a country, Census of import and export, etc.

Census investigation is useful in following situations:

1. When a deep study to be performed;
2. When study area is limited;
3. When an adequate accuracy and reliability is desired;
4. When investigator have resources; and
5. When use of sampling method is tough and prohibited.

1.3.1 Merits and Demerits of Census

Merits

1. **Useful in Heterogeneity**
   This method is very appropriate when the units are heterogeneous from each other and hard to be succeeded for sampling method.

2. **Deep Study is Possible**
   Through census, deep study of the subject is possible so that the investigator can get the total information of the variable of interest. He also knows the things or subjects which otherwise overlooked.

3. **High Level of Accuracy**
   High level of accuracy is expected by following this method. Because of that the personnel investigation is to be performed in investigation area therefore the results are accurate at higher degree.

4. **Necessary in Some Situation**
   If the nature of investigation is like that the involvement of all units is necessary than census is necessary i.e. census of population of a country.

Demerits

1. **Useless in Case of Destructive Units**
   If the units are destructive type and been destroyed by examining, this method is useless. For example, to study about the hardness of a chawk or quality of crackers or life of bulbs or tube lights.

2. **More Time and Energy**
   This method is very time consuming and large number of persons required to complete the process. This method consumes much of energy and hard work to perform the study.

3. **More Expensive**
   Much of time and organization of big size is needed for a census investigation. A large number of field investigators to be involved in this

work and arrangement of their training is also needed. In all, in this process large scale expenditure is needed.

4. **Investigation Remains Incomplete**
   In this method time, money, organization and large number of field investigators are required. Population is also large. Therefore, the investigation may remain incomplete due to weakness of investigating team or time or lack of availability of resources. In that situation, the effort or expenses which have already done that become useless.

5. **Inconvenient**
   This method is very inconvenient because this needs a whole department to be established separately. Problems related to management arise.

6. **Not Possible for all**
   This method is not comfortable for all because this method can be used solely by powerful person or organizations.

7. **Not Possible in Every Situation**
   Census is not possible from many reasons in various situations, where the investigation area is large and wide. In such cases contact to each and every unit is not possible.

8. **Statistical Error**
   In this method we cannot have the knowledge of statistical error.

E 2) Describe the census and situations where it is essential.

1.4 SAMPLE SURVEY

A finite subset of statistical individuals in a population is called a sample and the number of individuals in a sample is called the sample size.

Sample is often used in our day to day practical life. For example, in a shop we assess the quality of rice, wheat or any other commodity by taking a handful of it from the bag and then decide to purchase it or not. A house wife normally tests the cooked food to find if they are properly cooked and contain the proper quantity of salt.

If the population is infinite, census is not possible. Also, if the units are destroyed in the course of inspection, 100% inspection though is not possible at all desirable. But even if the population is finite or the inspection is not destructive, 100% inspection is not taken recourse too, because of the administrative, financial and time factor related problems. So we take the help of sampling.

1.4.1 Merits and Demerits of Sample Survey

Merits

1. **Detailed Inspection**
   As selected information is large so their detailed inspection can be done.

2. **Statistical Error**
   The investigator can analyse the statistical error from only the size of the sample in their investigation.

3. **Good Representative**
   If the proper selection is done than the result will be as same as after a census.

4. **Easy and Less Expensive**
   This method is easy and less expensive. It saves time, money and energy.

5. **Appropriate for Social and Economic Problem**
   As this method takes less time, therefore, this method is very appropriate for fast changing social and economic problems.

6. **Scientific**
   This method is more scientific because study may be done by other samples for available information.

Demerits

1. **Possibility of Inaccurate Result**
   If an investigator is biased at the time of selection of sample units then the result would be inaccurate.

2. **Inappropriate in Lack of Homogeneity**
   Where the lack of homogeneity exists or every unit being different type and nature this method cannot be adopted.

3. **Inappropriate in High Level Accuracy**
   This method is not appropriate in case where high level of accuracy is needed.

4. **Confused Result**
   If a suitable sampling method is not adopted or the sample size is not sufficient then the results would be incorrect.

1.4.2 Conditions for Sample Survey

The investigations through sample survey are appropriate in the following conditions:

1. **Broad Area**
   When the investigation area is broad, for example, testing of the effect of a drug for disease by a drug company, then they have to adopt the sample survey method.

2. **Implication of Rules**
   When implications of rules are to be done in a broad way, use of this method is advisable because conformation of rules can be done by various samples.

3. **When the Population is Infinite**
   When the number of elements in the population is infinite then this method is suitable i.e. counting of leaves in a tree is a tough job.

4. **Insufficient Resources**
   Where the money, time and employee/workers are in insufficient numbers, then this method can be adopted.

5. **No Need of High Level of Accuracy**
   Where a very high level of accuracy is not necessary, then this method can be used.

6. **When Units are of Destructive Nature**
   In some situations, if the units of the population are of destructive nature and if the census method is used then all the population would be destroyed. In this situation, the sample survey is advisable, for example, testing of sound of crackers.

7. **Use of Census is not Advisable as well as not Possible**
   Some investigation situations, where census is not only inappropriate but impossible also the sample survey is appropriate. For example if it has to know that in India's coal mine, how much and which type of coal existed so for that the sample survey is appropriate.

8. **Homogeneity**
   If the elements of a population are homogeneous than sample units would be of same characteristics as of the population. In short a sample should be a true representative of population.

E 3) Define a sample and describe the conditions for sample survey briefly.

1.5 PRINCIPLES OF SAMPLE SURVEY

Three basic principles for the design of a sample survey are:

1. **Principle of Optimization**
   The principle of optimization takes into account the factors of
   (a) Efficiency and (b) cost.

   (a) **Efficiency**
   Efficiency is measured by the inverse of sampling variance of the estimator. The principle of optimization ensures that a given level of efficiency will be reached with the minimum possible resources and minimum cost.

   (b) **Cost**
   Cost is measured by expenditure incurred in terms of money or man powers. So, the term optimization means that, it is based on developing methods of sample selection and of estimation; these provide a given value of cost with the maximum possible efficiency.

2. **Principle of Validity**
   By validity of a sample design, we mean that the sample should be so selected that the results could be interpreted objectively in terms of probability. According to this, sampling provides valid estimates about population parameters. This principle ensures that there is some definite

and preassigned probability for each individual of the aggregate (population) to be included in the sample.

3. **Principle of Statistical Regularity**
   According to the principle of statistical regularity we mean that 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. This principle has also its origin in the law of large numbers of the theory of probability.

1.5.1 Essentials of Sampling

For obtaining the unbiased and real result by a sampling method, a sample should have the following factors (characteristics):

1. **Homogeneity**
   The nature of each and every unit of the population should not contain much difference. If two or more samples are selected then they should be similar in nature not in their response/output.

2. **Representativeness**
   The sample should represent all the characteristics of the population that can be possible only when the selection of items or units has been done unbiased and each and every unit have an equal probability of chance to be selected in the sample.

3. **Independency**
   Each and every unit of the population should be independent. In other words, the selection of a unit in the sample should not be dependent on the selection of other units.

4. **Adequacy**
   The number of units or elements which are to be selected in the sample should be sufficient. If the sample size is not sufficient then results cannot be reliable. The more the sample units in the sample, more reliable results would occur.

E 4) What do you mean by the principle of optimization?

1.6 PRINCIPLE STEPS IN SAMPLE SURVEY

The main steps involved in the planning and execution of a sample survey are under the following heads:

1. **Objectives**
   The objective of the survey must be defined in clear and concrete terms. Generally, in survey a investigation team is not quite clear in mind as to what they want and how they are going to use the results. Some of the objectives may be immediate and some far-reaching. The investigator should take care of these objectives with the available resources in terms of money, manpower and the time limit required for the availability of the survey.

2. **Defining the Population**
   The population from which sample is chosen should be defined in clear and unambiguous terms. The geographical, demographic and other boundaries of the population must be specified so that no ambiguity arises regarding the coverage of the survey.

3. **Sampling Frame and Sampling Units**

   The sampling unit is the ultimate unit to be sampled for the purpose of the survey. The sampling units must cover the entire population and they must be distinct, unambiguous and non-overlapping in the sense that every element of the population belongs to one and only one sampling unit. In a Socio economic survey, whether a family or a member of a family is to be the ultimate sampling unit.

   Once the sampling units are defined, one must see whether a sampling frame which is a list of all the units in the population, is available. The construction of the frame is often one of the major practical problem since it is the frame which determines the structure of the sample survey. The list of units have to be carefully scrutinized and examined to ensures that it is free from duplicity or incompleteness and are up-to-date. A good frame is hard to come by and only good experience helps to construct a good frame.

4. **Selection of Proper Sampling Design**
   This is the most important step in planning a sample survey. There is a group of sampling designs (to be discussed later) and selection of the proper one is an important task. The design should take into account the available resources and the time-limit, if any, besides the degree of accuracy desired. The cost and precision should also be considered before the final selection of sampling design.

5. **Method of Collecton of Data**
   For collection of data, either the interview method or the mail questionnaire method is to be adopted. Although the later method is less costly but there is a large scope of non-response in it. In the cases, where the information is to be collected by observation they must decide upon the method of measurement.

6. **Data to be Collected**
   Collection of data must be done in conformity with the objectives of the survey and the nature of the data. After it is decided upon, one must prepare a questionnaire or a schedule of enquiry. A schedule or a questionnaire contains a list of items of which information is sought, but the exact form of the questions to be asked is not standardized but left to the judgment of the investigators. A questionnaire should be in a specified order. The questions should be clear, brief, collaborative, non offending and unambiguous and to the point so that not much scope of speculation is left on the part of the respondent or interviewer.

7. **Field Work Organization**
   Field work, itself has several stages and so it is to be well organized. The different stages includetraining the field workers, supervising the field workers, etc. It is absolutely essential that the personnel should be

thoroughly trained in locating the sample units, the methods of collection of required data before starting the field word. The success of a survey to a great extent depends upon the reliable field work. Inspection after field work by the adequate supervisors should also be performed.

8. **Summary and Analysis of Data**
   This is the last step wherein inference is to be made on the basis of collected data. This step again consists of the following steps:

   a) The filled in questionnaires should be carefully scrutinized to find out whether the data furnished are plausible and consistent;
   b) Depending upon the quantity of data, a hand-tabulation or machine tabulation is to be drawn;
   c) After the data has been properly scrutinized, edited and tabulated, a very careful statistical analysis is to be made; and
   d) Finally a report incorporating detailed statement of the different stages of the survey should be prepared. In the presentation of the result, it is advisable to report technical aspects of the design.

Let us answer the given exercise.

E 5) List the principles steps of sample surveys.

1.7 SAMPLING AND NON-SAMPLING ERROR

The errors involved in the collection, processing and analysis of data may be broadly classified under the following two heads:

1. Sampling Error
2. Non-sampling Error

1.7.1 Sampling Error

The error which arises only in sample survey is termed as sampling error. This error arises because in sample survey a part of the population is only studied. This is the reason why sampling error is absent in census. The main factors of sampling error are:

1. Some of the bias is introduced by the use of defective sampling techniques for the selection of a sample;
2. Substitution of a non-selected a convenient unit of the population in place of a selected unit to which the investigation is difficult leads to some biases in the sample survey;
3. Bias due to defective demarcation of sampling units, particularly in area/filed survey; and
4. Constant errors due to improper choice of the statistics for estimating the population parameters.

1.7.2 Non-Sampling Error

The non-sampling error arises at the stages of observation, ascertainment and processing of the data. This is the reason why the non-sampling error presents

in both the census and the sample survey. Non-sampling error can occur at every stage of the planning or execution of census or sample survey. Non-sampling errors arise due to the following factors:

1. Data specification being inadequate and inconsistent with respect to the objective of the study;
2. Error due to location of the units and actual measurement of the characteristics;
3. Error due to ill designed questionnaire;
4. Lack of trained and qualified investigators and lack of adequate supervisory staff;
5. Errors due to lack of correct responses furnished by the respondents;
6. Non-response biases occur if full information is not obtained on all the sampling units;
7. If the objectives of the survey are not stated clearly, it may result in inclusion of the units which are not to be included and exclusion of the units which are to be included in the sample;
8. Due to error in various operations of data processing such as editing and coding of the responses, punching of cards, tabulation and summarizing the observation made in the survey; and
9. The errors may be committed during presentation and printing the results of the survey.

Let us answer the given exercise.

E 6) Distinguish between the sampling and non-sampling Error.

1.8 ADVANTAGES OF SAMPLING OVER CENSUS

The advantages of sampling over complete census may be outlined as follows:

1. Sampling requires less time and labor than census because only a part of the population has to be examined. The sampling results also can be analysed much faster;
2. Sampling usually results in reduction in cost in terms of money and man powers. The total cost of the sample survey is expected to be much smaller than a complete census;
3. There is generally a greater scope in a sample survey than in census. Some inquiries may require highly trained personnel or specialized equipment for collection of data, then the census may be inconceivable;
4. In some cases a complete census is ruled out by the nature of the population. If there is a population which is infinite and/or hypothetical, then sampling is the only option;
5. A sample survey gives data of better quality than a complete census, because in a sample survey it may be possible to use better resources than complete census;

6. If the population is too large, as for example, trees in a jungle, leaves in a tree i.e. we are left with no option but to resort to sampling; and

7. If testing is destructive, then complete enumeration is impracticable and sampling design is the only method to be used in such cases. For example, testing the breaking strength of a chalk, testing of lifetime of an electrical bulb, etc.

Let us answer the given exercise.

E 7) Describe the advantages of sampling over census.

1.9 TYPES OF SAMPLING

According to the method of selection of sample, the sampling schemes can be categorised as follows:

1. Non-probability sampling;
2. Probability or random sampling; and
3. Mixed sampling.

1.9.1 Non-Probability Sampling

In this method, the sample is selected with a definite purpose in view and the choice of the sampling units depends entirely on the discretion and judgment of the investigator. While selecting a sample, investigator tries to include each and every characteristics of population in sample.

Non-Probability Sampling scheme can be classified as:

1. **Purposive Sampling**
   In this sampling the sample is selected with definite purpose in view and the choice of sampling units depends entirely on the discretion of the surveyor. This sampling suffers from drawback of favoritism and nepotism of the surveyor.

2. **Judgment Sampling**
   In judgment sampling respondents are selected on the judgment of the surveyor with the hope that they will meet requirements of the study. The underling assumptions are that the respondent truly represents the entire population. To find out the potential guide for the food and catering technology a researcher go to the teachers of Hotel Management Department may be the example of judgment sampling.

3. **Deliberate Sampling**
   In deliberate sampling, deliberate selection of sample is made so that any important unit could not be leftout.

4. **Convenience Sampling**
   In convenience sampling method, a surveyor selects the sample at his/her own convenience, often as the study is being conducted. Convenience

sampling is based on assumption that the target population is homogeneous and the individuals selected and interviewed yields similar information with regard to the characteristics under study. If persons selected from restaurants to collect the information about quality of the food, service, etc. are supposed to represent the population of food takers. Such a sampling is known as convenience sampling.

5. **Quota Sampling**
   If the cost of selected random samples in each stratum is very high in stratified sampling then the sampling units are assigned in a quota (fixed number of units) in each stratum and the actual selection of units is left at the decision of the surveyor.

Merits of Non-Probability Sampling

1. This method of sampling is very simple;
2. After sample size determination with the help of planning, a suitable sample may possibly be obtained; and
3. Important units or members may be included in the sample.

Demerits of Non-Probability Sampling

1. Predetermined view of selector effects the selection of sample which impure the result. This effect does directly or indirectly on the process;
2. There is no place for probability in selection of units therefore sampling error cannot be obtained;
3. There is no guaranty of validity of the results from the sample selected by this method; and
4. The attitude and biasedness of the investigator also affect the selection of sample that's why the results obtained by this method are not reliable scientifically.

Significance of Non-Probability Sampling

1. When the number of units in population is less and there is possibility that the important units may be left;
2. When sample size is to be kept small;
3. When the deep study of the important unit (purposive unit) is to be done;
4. When the investigator has the experience of obtaining a correct sample; and
5. This method can be used in pilot survey.

1.9.2 Random or Probability Sampling

The technique of random sampling is of fundamental importance in the application of Statistics. The estimation theory is based on the assumption of random sampling. Probability sampling is the scientific method of selecting samples accordingly to some laws of chance in which each unit in the population has some definite pre-assigned probability of being selected in the sample. The following are the different types of probability sampling:

1. Where each unit has an equal chance of being selected;
2. Sampling units have different probabilities of being selected; and

3. Probability of selection of a unit is proportional to the sample size.

Merits of Random or Probability Sampling

1. **No Plan for Selection**
   There is no need to make any detailed plan for the selection of units.

2. **Less Expensive**
   In this method, money, time and hard work are very less.

3. **Unbiased**
   In this method there is no space for any biasedness. Every unit has same chance of selection.

4. **Inspection of Purity**
   Inspection of purity of one sample can be done by other sample. In this method measure of statistical error can also be done.

5. **Random Selection**
   Selector has not to use his mind. He selects units at random.

6. **True Representation of Population**
   In this method real characteristics can be represented through sample because it is based on the law of statistical regularity and law of inertia. In real, it becomes a small part of population.

Demerits of Random or Probability Sampling

1. **Inappropriate**
   This method is not appropriate where some units are so important to be included necessarily in the sample.

2. **Less Representative**
   It may be possible that sample could not represent the population if sample is not sufficiently large.

3. **Less Independency**
   This method is useless if the units of the population are dependent.

Limitations of Random Sampling

1. If investigation area is small then results may not be reliable;
2. If all the units are heterogeneous then sample cannot be true representative;
3. The elements of the population should be independent; and
4. The results would not be reliable if the investigator is biased.

1.9.3 Difference between Probability and Non-probability Sampling

1. In non-probability sampling the selection of units are pre-decided whether in probability sampling is based on chances.
2. Non-probability sampling is biased but probability sampling is unbiased.
3. In non-probability sampling, the errors are of cumulative in nature whereas in probability sampling errors are less.

4. If a sample, from a population with homogeneous and important units, is to be selected then non-probability sampling is appropriate where as probability sampling is used in various kinds of population.

1.9.4 Mixed Sampling

If the samples are selected partly according to some laws of chance and partly according to a fixed rule, they are called mixed samples and the method of selecting such samples is known as mixed sampling. The merits of this sampling are the mixture of the merits of both sampling. Selection of units is more reliable in this method because that is the representation of various stages of population. In mixed sample no important characteristics is left which is to be selected in the sample.

E 8) Describe the types of sampling briefly.

1.10 OBJECTIVES OF SAMPLING

Sampling investigation can be performed to fill the following objectives:

1. **Checking of Validity of Census**
   Validity of results obtained from census investigation is to be checked by sampling. To check the validity of the census results the sample survey are to be organized after census.

2. **Checking of Difference between the Measurements of a Sample and Population**
   There is always a difference between the measurements of population and its sample whether the sample is suitable enough. Even, there is always difference between the measures of two samples of the same population. Inspection of authenticity of these differences is the main objective of sampling.

3. **Checking of Characteristics of Population**
   This is the main objective of a sampling study that all the characteristics of the population can be found in less time, through less effort and with least cost. More information can be obtained about the whole population through sampling.

4. **Find Estimate of Parameters of Population**
   The measure of the population is to be obtained on the basis of statistical measures of sample mean; sample standard deviation, sample correlation, etc. In this way the objective of the sampling is to find out the most probable values of the parameters. The measures of the population are called parameters and the measures of sample are called statistic.

5. **Fulfillment of Special Objective and for Continuous Information**
   Sampling methods are applied to fulfill the specific objective. Social, economic and behavioral surveys come under this. Continuous information about the behavior of the unit is needed for some population. Through sampling, corrections are to be done regularly in the results of the previous sample survey as in Medical Sciences and Quality Control.

1.11 PROBLEMS OF SAMPLING METHODS

Some problems arise regarding the sampling on which pre discussion is necessary:

1. **Determination of Base of Sampling**

   Sample should neither be too small nor too big. This is to be noted that a sample is surely would not be a representative only by size. Sample size depends on the following points:

   1. Homogeneity or heterogeneity of population;
   2. Nature of investigation;
   3. Practical things like money, time, hard work by trained supervisors, etc.;
   4. Level of purity; and
   5. Sampling method.

2. **Discussion on Effect of Biasedness in Sampling**
   Partiality, predecided and unknowingly happenings make impure the sample. So it is necessary to beware of that.

3. **Reliability in Sampling**
   This can be done in following ways:
   1. Comparison of sample after dividing in two equal parts;
   2. Selection of another sample of same size from the same population and comparison; and
   3. Comparison of the result of a sub-sample to the result of the sample itself.

1.12 SUMMARY

In this unit, we have discussed:

1. A population and the different kinds of populations;
2. The census method and sampling method and comparison of both;
3. The conditions and principles of sample surveys;
4. The principle steps in sample survey;
5. The sampling and non-sampling error;
6. The different kinds of sampling procedure; and
7. The objectives, problems and importance of sampling.

1.13 SOLUTIONS /ANSWERS

E1) Same as the Section 1.2 and Sub-section 1.2.1

E2) In census we study about each and every unit of the population, that means whole group related to investigation is investigated and the information are to be collected. Population means total units of

investigation area i.e. census of population, census of import and exports, etc.

Census investigation is useful in following situation

1. When a deep study to be performed;
2. When study area is limited;
3. When an adequate accuracy and reliability is desired;
4. When investigator have resources; and
5. When use of sampling method is tough and prohibited.

E3) A finite subset of statistical individuals in a population is called a sample and the number of individuals in a sample is called the sample size. Sample is oftenly used in our day to day practical life. For example, in a shop we assess the quality of rice, wheat or any other commodity by taking a handful of it from the bag and then decide to purchase it or not. Rest part of the answer is as same as Sub-section 1.4.2.

E4) The principle of optimization takes into account the factors of

(a) Efficiency and (b) Cost.

Efficiency is measured by the inverse of sampling variance of the estimator. The principle of optimization ensures that a given level of efficiency will be reached with the minimum possible resources and will be attained with available resources and cost.

Cost is measured by expenditure incurred in terms of money or manpower's. The term optimization means that, it is based on developing methods of sample selection and of estimation; these provide a given value of cost with the maximum possible efficiency.

E5) The main steps involved in the planning and execution of a sample survey are as follows:

1. Defining the population
2. Defining the objectives of the sample survey
3. Method of data collection
4. Selection of proper sampling design
5. The frame and the sampling unit
6. Data collection
7. Field work organization
8. Analysis of data and summary

E6) The error which arises only in sample survey is termed as sampling error. This error arises because a part of the population is only studied. This is the reason why sampling error is absent in census. The main factors of sampling error are:

1. Some of the bias is introduced by the use of defective sampling techniques for the selection of a sample;
2. Substitution of a non-selected convenient unit of the population, in place of a selected unit to which investigation is difficult, leads to some biases in the sample survey;

3. Bias due to defective demarcation of sampling units is particularly in area survey; and
4. Constant errors due to improper choice of the statistics for estimating the population parameters.

The non-sampling errors arise at the stages of observation, ascertainment and processing of the data. This is the reason why the non-sampling error presents in both the census and the sample survey. Non Sampling error can occur at every stage of the planning or execution of census or sample survey. Non-Sampling errors arise from the following factors:

1. Data specification being inadequate and inconsistent with respect to the objective of the study;
2. Error due to location of the units and actual measurement of the characteristics, error due to ill designed questionnaire, etc.;
3. Lack of trained and qualified investigators and adequate supervisory staff;
4. Errors due to lack of correct responses furnished by the respondents;
5. Non-response biases occur if full information is not obtained on all the sampling units;
6. If the objectives of the survey are not stated clearly that may result in inclusion of the units which are not to be included and execution of the units which are to be included in the sample;
7. Due to error in various operations of data processing such as editing and coding of the responses, punching of cards, tabulation and summarizing the observation made in the survey; and
8. The errors may be committed during presentation and printing the results of the survey.

E7) Same as Section 1.8

E8) According to the method of selection of sample, the sampling schemes can be categories as follows:

1. Non-Probability Sampling;
2. Probability or Random Sampling; and
3. Mixed Sampling.

1. **Non-Probability Sampling**

In this method, the sample is selected with definite purpose in view and the choice of the sampling units depends entirely on the discretion and judgment of the investigator. While selecting a sample investigator tries to include each and every characteristics of population in sample.

Non-Probability Sampling scheme can be classified as:

1. Purposive Sampling
2. Judgment Sampling
3. Deliberate Sampling
4. Convenience sampling
5. Quota Sampling

2. **Random or Probability Sampling**

The technique of random sampling is of fundamental importance in the application of Statistics. The estimation theory is based on the assumption of random sampling. Probability sampling is the scientific method of selecting samples according to some laws of chance in which each unit in the population has some definite pre assigned probability of being selected in the sample. The following are the different types of probability sampling:

1. Where each unit has an equal chance of being selected;
2. Sampling units have different probabilities of being selected; and
3. Probability of selection of a unit is proportional to the sample size.

Some of the Probability Sampling schemes are:

1. Simple Random sampling;
2. Stratified Random sampling;
3. Systematic Random sampling; and
4. Cluster sampling, etc.

3. **Mixed Sampling**

If the samples are selected partly according to some laws of chance and partly according to a fixed rule, they are called mixed samples and the method of selection of such samples is known as mixed sampling. The merits of this sampling are the mixture of the merits of both sampling. Selection of units is more reliable in this method because that is the representation of various stages of population. In mixed sample no important characteristics is left which is to be selected in the sample.

UNIT 2 SIMPLE RANDOM SAMPLING

Structure

2.1 Introduction
Objectives
2.2 Methods of Selection of a Sample
Lottery Method
Random Number Method
Computer Random Number Generation Method
2.3 Properties of Simple Random Sampling
Merits and Demerits of Simple Random Sampling
2.4 Simple Random Sampling of Attributes
2.5 Sample Size for Specific Precision
2.6 Summary
2.7 Solutions/Answers

2.1 INTRODUCTION

Simple random sampling refers to the sampling technique in which each and every item of the population is having an equal chance of being included in the sample. The selection is thus free from any personal bias because the investigator does not make any preference in the choice of items. Since selection of items in the sample, depends entirely on chance, this method is also known as the method of probability sampling.

Random sampling is sometimes referred to a representative sampling. If the sample is chosen at random and the size of the sample is sufficiently large, it will represent all groups in the population. An element of randomness is necessary to be introduced in the final selection of the item. If that is not introduced, bias is likely to enter and make the sample unrepresentative.

Methods of selection of a simple random sample are explained in Section 2.2. In Section 2.3 the properties of simple random sampling are described. Simple random sampling of attributes is introduced in Section 2.4 whereas in Section 2.5 the sample size determination for specific precision is described briefly.

Objectives

After studying this unit, you would be able to

- describe the simple random sampling;
- explain the method of SRSWR and SRSWOR;
- explain and derive the properties of simple random sampling;
- calculate the variance of the estimate of the sample mean;
- describe the SRS for attribute and its properties; and
- describe and draw a appropriate sample size for specific precision.

2.2 METHODS OF SELECTION OF A SAMPLE

The random sample obtained by a method of selection in which every item has an equal chance to be selected in the sample. The random sample depends not only on method of selection but also on the sample size and nature of population.

Simple Random Sampling
If a sample of n units is selected randomly from a population of size N, this method is known as simple random sampling. As the name suggest, simple random sampling is a method in which the required number of elements /units are selected simply by random method from the target population. One can select a simple random sample by either of these two methods with replacement method and without replacement method.

Simple Random Sampling with Replacement (SRSWR)
When simple random samples are selected in the way that units which has been selected as sample unit is remixed or replaced in the population before the selection of the next unit in the sample then the method is known as simple random sampling with replacement.

Simple Random Sampling without Replacement (SRSWOR)
When simple random sample are selected in the way that a unit is selected as sample unit is not mixed or replaced in the population before the selection of the next unit. This method is known as simple random sampling without replacement i.e. once a unit is selected in the sample will never be selected again in the sample.

Some procedures are simple for small population and is not so for the large population. Proper care has to be taken to ensure that selected sample is random. Random sample can be obtained by any of the following methods:

1. By Lottery method;
2. By 'Mechanical Randomization' or 'Random Numbers' method; and
3. By Computer Random Number Generation method.

2.2.1 Lottery Method

This is very popular method of selecting a random sample under which all items of the population are numbered or named on separate slips of paper. These slips of paper should be of identical size, color and shape. These slips are then folded and mixed up in a container or box or drum. A blind fold selection is then made of the number of slips required to constitute the desired size of sample. The selection of items thus depends entirely on chance. For example if we want to select n candidates out of N. We assign the numbers 1 to N. One number to each candidate and write these numbers (1 to N) on N slips which are made as homogeneous as possible. These slips are then put in bag and thoroughly shuffled and then n slips are drawn one by one. Then the n candidates corresponding to numbers on the slip drawn will constitute a random sample.

If we draw a slip and note down the number written on the slip and then again replace the slip in the bag and then draw the next. These number of slips constitute a sample of required size is called a sample of SRSWR. If we do not replace the first slip which has already been drawn in the bag for the subsequent draws then it is called SRSWOR.

The above method is very popular in lottery draws where a decision about prizes is to be made. However, while adopting lottery method, it is absolutely essential to see that the slips are as homogeneous as possible in terms of size, shape and color, otherwise there is a lot of possibility of personal prejudice and bias affects in the result.

2.2.2 Random Number Method

The lottery method, discussed above, become quite cumbersome to use if the size of the population is very large. An alternative method of random selection is that of using the table of random numbers. The most practical and inexpensive method of selecting a random sample consists in the use of 'Random number tables' which have been so constructed that each of the digits 0, 1, 2, … , 9 appears with approximately the same frequency and independently.

If we have to select a sample from a population of size N ($\leq$ 99) then the numbers can be combined two by two to give pairs from 00 to 99. The method of drawing the random sample consists in the following steps:

1. Identify the N units in the population with the numbers from 1 to N;
2. Select at random, any page of the random number tables and pick up the numbers in any row or column or diagonal at random and discart the number which is greater than N; and
3. The population units corresponding to the numbers selected in step-2, constitute the random sample.

The sample may be selected with replacement or without replacement. In the case of sampling with replacement, a number occurring more than once is accepted. A unit is repeated as many times as a random number occurs. But in the case of sampling without replacement, if a number in random number table or remainder occurs more than once is omitted at any sequent stage. In the above selection procedure numbering of units from 00 onwards and making use of remainders has an advantage, as no random numbers is being wasted during the selection procedure. This saves time and labor. For example a population consists of 20 units and a sample of size 5 is to be selected from this population. Since 20 is a two digit figure, unit are numbered as 00, 01… 19. Five random numbers are obtained from a two digit random number table. They are given as 85, 63, 52, 34, and 46. On dividing 85 by 20 the remainder is 5, hence select the unit on serial no. 5. Similarly, dividing 63, 52, 34 and 46 by 20, the respectively remainder are 3, 12, 14 and 6. Hence selected units are at serial numbers 05, 03, 12, 14 and 06. These selected units constitute the sample.

A number of random number tables are available such as:

1. **Tippet's Random Number Tables**
   These tables consist of 10,400 four digits numbers, giving in all 10,400 $\times$ 4 i.e. 41,600 digits.

2. **Fisher and Yate's Random Number Tables**
   These tables consist of 15000 digits arranged in two digits numbers.

3. **Kendall and Babington Smith's Random Tables**
   These tables consist of 1,00,000 digits grouped into 25,000 sets of 4 digits random numbers.

4. **Rand Corporation Random Number Tables**
   These tables consist of one million random digits consisting of 2,00,000 random numbers of 5 digits each.

2.2.3 Computer Random Number Generation Method

The main disadvantage of random number method is that if we want to draw a sample of large size from the target population then it will take a long time to draw random numbers from the random number table. Thus to save time and energy in generation of large numbers of a random numbers one can opt computer random numbers generation method. Many kinds of methods have been used for generation of random number from computer but we shall not discuss all of them here and describe only linear congruential generation method.

Linear Congruential Generation Method

Linear congruential method can take many different forms but the most commonly used form is defined by

$z_i = (az_{i-1} + c) \bmod m \quad \text{for } i = 1, 2, \ldots$

where, a, c and z_i are to be in the range $(0, 1, 2, \ldots, m-1)$ and integers

a - Multiplier Integer

c - Shift or Increment Integer

m - Modulus

Here, (n) mod m means remainder term when n is divided by m.

Suppose we want to draw a sample of size 4 from the population of size 8.

Take m = 8, a = 5, c =7 and $z_0 = 4$ the resulting sequence of random numbers is calculated as

$z_1 = (5 \times 4 + 7) \bmod 8 = 3$

$z_2 = (5 \times 3 + 7) \bmod 8 = 6$

$z_3 = (5 \times 6 + 7) \bmod 8 = 5$

$z_4 = (5 \times 5 + 7) \bmod 8 = 0$

For above calculation we may also make a computer program which is beyond this course. We are, therefore, not going to discuss it here.

2.3 PROPERTIES OF SIMPLE RANDOM SAMPLING

Terminologies

N = Population size

n = Sample size

X_i = Value of the character under study for the i^{th} unit in the population

x_i = Value of the character under study for the i^{th} unit in the sample

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \text{Population mean}$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \text{Sample mean}$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 = \text{Population mean square}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \text{Sample mean square}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 = \text{Population variance}$$

Theorem 1: Prove that the probability of selecting a specified unit of the population at any given draw is equal to the probability of its being selected at the first draw.

Proof: In simple random sampling method an equal probability of selection is assigned to each unit of the population at the first draw.

Thus, in SRS from a population of N units, the probability of drawing any unit at the first draw is $\frac{1}{N}$, the probability of drawing any unit in the second draw from among the available $(N-1)$ is $1/(N-1)$ and so on.

Let, E_r be the event that any specified unit is selected at the r^{th} draw.

$P(E_r)$ = Prob.{A specific unit is not selected at any one of previous $(r-1)$ draws and then selected at the r^{th} draw}

$$= \sum_{i=1}^{r-1} P(\text{It is not selected at } i^{th} \text{ draw}) \times P(\text{It is selected at } r^{th} \text{ draw that is not selected at the previous } (r-1) \text{ draw})$$

$$P(E_r) = \sum_{i=1}^{r-1}\left[1 - \frac{1}{N-(i-1)}\right] \times \frac{1}{N-(r-1)}$$

$$P(E_r) = \sum_{i=1}^{r-1}\left[\frac{N-i}{N-(i-1)}\right] \times \frac{1}{N-(r-1)}$$

$$P(E_r) = \frac{(N-1)}{N} \times \frac{(N-2)}{(N-1)} \times \frac{(N-3)}{(N-2)} \times \ldots \times \frac{(N-r+1)}{(N-r+2)} \times \frac{1}{(N-r+1)}$$

$$P(E_r) = \frac{1}{N}$$

That means

$$P(E_r) = \frac{1}{N} = P(E_1)$$

Theorem 2: The probability that a specified unit is selected in the sample of size n is $\frac{n}{N}$

Proof: Since a specified unit can be selected in the sample of size n in n mutually exclusive ways, viz. it can be selected in the sample at the r^{th} draw (r =1, 2,…, n) and since the probability that it is selected at r^{th} draw is

$$P(E_r) = \frac{1}{N} \qquad ; r = 1, 2, 3, ..., n$$

Therefore, the probability that a specified unit is included in the sample would be the sum of the probabilities of inclusion in the sample at 1^{st} draw, 2^{nd} draw, …, n^{th} draw. Thus, by addition theorem of probability, we get

$$P\left(\bigcup_{r=1}^{n} E_r\right) = \sum_{r=1}^{n} \frac{1}{N} = \frac{n}{N}$$

Theorem 3: The possible numbers of sample of size n from a population of size N if sampling is done with replacement is N^n.

Proof: The first unit can be drawn from N units in N ways. Similarly, second unit can also be drawn in N ways because the first selected unit again mixed with the population. So on up to the selection of n^{th} unit.

Thus, the total number of ways are

$${}^{N}C_1 . {}^{N}C_1 . {}^{N}C_1 ... {}^{N}C_1 \ (n \text{ times})$$

$$\Rightarrow \left({}^{N}C_1\right)^n = N^n$$

Theorem 4: In SRSWOR the sample mean $\bar{x}$ is an unbiased estimator of population mean $\bar{X}$.

Proof: We have

$$E(\bar{x}) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = E\left[\frac{1}{n}\sum_{i=1}^{N} a_i X_i\right]$$

where, a_i = 1 if i^{th} unit is included in the sample
0 if i^{th} unit is not included in the sample

Since, a_i takes only two values 1 and 0

$$E(a_i) = 1.P(a_i = 1) + 0.P(a_i = 0)$$

= 1.P (i^{th} unit is included in a sample of size n) + 0.P (i^{th} unit is not included in the sample)

$$= 1.\frac{n}{N} + 0.\left(1 - \frac{n}{N}\right) = \frac{n}{N}$$

Hence,

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

Theorem 5: In SRSWR, the sample mean $\bar{x}$ is an unbiased estimator of population mean $\bar{X}$.

Proof: We have

$$E(\bar{x}) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]$$

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

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

$$= \frac{1}{n} n.\bar{X} = \bar{X}$$

Theorem 6: In SRSWOR, the sample mean square is an unbiased estimate of the population mean square, i.e.

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

Proof: We have

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

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i\neq j=1}^{n} x_i x_j\right]$$

$$= \frac{1}{n-1}\left[\left(1-\frac{1}{n}\right)\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i\neq j=1}^{n} x_i x_j\right]$$

$$= \frac{1}{n-1}\times\frac{(n-1)}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} x_i x_j$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} x_i x_j$$

$$E(s)^2 = \frac{1}{n}E\left[\sum_{i=1}^{n} x_i^2\right] - \frac{1}{n(n-1)}E\left[\sum_{i\neq j=1}^{n} x_i x_j\right] \quad \ldots (1)$$

We have

$$E\left[\sum_{i=1}^{n} x_1^2\right] = E\left[\sum_{i=1}^{N} a_i X_i^2\right]$$

$$= \sum_{i=1}^{N} E(a_i) X_i^2 \quad \ldots (2)$$

where, a_i = 1 if i^{th} unit is included in the sample

0 if i^{th} unit is not included in the sample … (3)

Therefore,

$$E\left[\sum_{i=1}^{n} x_i^2\right] = \frac{n}{N}\sum_{i=1}^{N} X_i^2 \quad \ldots (4)$$

and

$$E\left[\sum_{i\neq j=1}^{n} x_i\, x_j\right] = E\left[\sum_{i\neq j=1}^{N} a_i a_j X_i\, X_j\right]$$

$$= \left[\sum_{i\neq j=1}^{N} E(a_i a_j) X_i\, X_j\right] \quad \ldots (5)$$

where, a_i and a_j are defined in equation (3)

Therefore,

$$E\left(a_i a_j\right) = 1.P\left(a_i a_j = 1\right) + 0.P\left(a_i a_j = 0\right)$$

$$\Rightarrow P\left[\left(a_i = 1\right) \cap \left(a_j = 1\right)\right]$$

$$\Rightarrow P\left(a_i = 1\right).P\left(a_{j=1} \,/\, a_i = 1\right)$$

$$\Rightarrow \frac{n(n-1)}{N(N-1)} \quad \ldots (6)$$

Because

$$E\left(a_i = 1\right) = P\left[i^{th} \text{ unit is included in the sample}\right] = \frac{n}{N}$$

and

$$P\left(a_j = 1 \,/\, a_i = 1\right) = P\begin{bmatrix} j^{th} \text{ unit is included in the sample given} \\ \text{that } i^{th} \text{ unit is included the sample} \end{bmatrix}$$

$$= \frac{n-1}{N-1}$$

Substituting in equation (5), we get

$$E\left[\sum_{i\neq j=1}^{n} x_i\, x_j\right] = \sum_{i\neq j=1}^{N} \frac{n(n-1)}{N(N-1)} X_i\, X_j \quad \ldots (7)$$

Substituting from equations (4) and (7) in equation (1), we get

$$E(s)^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \frac{1}{N(N-1)}\sum_{i\neq j=1}^{N} X_i X_j$$

$$= \frac{1}{N-1}\left[\frac{N-1}{N}\sum_{i=1}^{N} X_i^2 - \frac{1}{N}\sum_{i\neq j=1}^{N} X_i X_j\right]$$

$$= \frac{1}{N-1}\left[\left(1-\frac{1}{N}\right)\sum_{i=1}^{N} X_i^{\,2} - \frac{1}{N}\sum_{i\neq j=1}^{N} X_i\, X_j\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N} X_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} X_i^2 + \sum_{i\neq j=1}^{N} X_i X_j\right)\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N} X_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} X_i\right)^2\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N} X_i^2 - N\overline{X}^2\right] = S^2$$

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

Theorem 7: In SRSWOR, the variance of the sample mean is given by

$$\mathrm{Var}(\bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

Proof: We have

$$\mathrm{Var}(\bar{x}) = E\left[\bar{x} - E(\bar{x})\right]^2$$

$$= E(\bar{x})^2 - \overline{X}^2 \qquad \ldots (8)$$

Now

$$E(\bar{x}^2) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^2$$

$$= \frac{1}{n^2} E\left[\sum_{i=1}^{n} x_i^2 + \sum_{i\neq j=1}^{n} x_i x_j\right]$$

$$= \frac{1}{n^2}\left[E\left(\sum_{i=1}^{n} x_i^2\right) + E\left(\sum_{i=1}^{n} x_j\right)\right] \qquad \ldots (9)$$

From equation (4), we have

$$E\left[\sum_{i=1}^{n} x_i^2\right] = \frac{n}{N}\sum_{i=1}^{n} x_i^2$$

But $$\sum_{i=1}^{N}(X_i - \overline{X})^2 = \sum_{i=1}^{N} X_i^2 - N\overline{X}^2$$

$$\Rightarrow \sum_{i=1}^{N} X_i^2 = \sum_{i=1}^{N}(X_i - \overline{X})^2 + N\overline{X}^2$$

$$\Rightarrow \sum_{i=1}^{N} X_i^2 = (N-1)S^2 + N\overline{X}^2$$

Therefore,

$$E\left(\sum_{i=1}^{n} x_i^2\right) = n\left[\left(\frac{N-1}{N}\right) S^2 + \overline{X}^2\right] \qquad \ldots (10)$$

Also from equation (7), we have

$$E\left(\sum_{i\neq j=1}^{n} x_i x_j\right) = \frac{n(n-1)}{N(N-1)}\sum_{i\neq j=1}^{N} X_i X_j$$

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

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

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

$$= n(n-1)\left[\overline{X}^2 - \frac{S^2}{N}\right] \qquad \ldots (11)$$

Substituting from equations (10) and (11) in equation (9), we get

$$E(\bar{x}^2) = \frac{1}{n}\left[\left(1-\frac{1}{N}\right)S^2 + \bar{X}^2\right] + \left(1-\frac{1}{n}\right)\left[\bar{X}^2 - \frac{S^2}{N}\right]$$

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

$$= \bar{X}^2 + \left(\frac{1}{n}-\frac{1}{N}\right)S^2 \quad \text{... (12)}$$

Substituting from equation (12) in equation (8), we get

$$Var(\bar{x}) = \bar{X}^2 + \left(\frac{1}{n}-\frac{1}{N}\right)S^2 - \bar{X}^2$$

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

Theorem 8: In SRSWR, variance of sample mean is given by

$$Var(\bar{x}) = \frac{(N-1)}{nN}S^2$$

Proof: We have

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

Again,

$$Var(\bar{x}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i)$$

Since in case of SRSWR each observation is independent, therefore

$$Var(\bar{x}) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2$$

$$= \frac{1}{n^2} n.\sigma^2 = \frac{1}{n}\sigma^2 \quad \text{… (13)}$$

But

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

$$\sigma^2 = \frac{N-1}{N}S^2 \quad \text{… (14)}$$

Substituting from equation (14) in equation (13) we get

$$Var(\bar{x}) = \frac{(N-1)}{nN}S^2$$

Theorem 9: The variance of the sample mean is more in SRSWR in comparison to its variance in SRSWOR, i .e.

$$Var_{SRSWR}(\bar{x}) > Var_{SRSWOR}(\bar{x})$$

Proof: We have

$$Var_{SRSWR}(\bar{x}) = \frac{N-1}{nN}S^2$$

and $$\text{Var}_{\text{SRSWOR}}(\bar{x}) = \frac{N-n}{nN}S^2$$

Therefore,

$$\text{Var}_{\text{SRSWR}}(\bar{x}) - \text{Var}_{\text{SRSWOR}}(\bar{x}) = \frac{(N-1)}{nN}S^2 - \frac{(N-n)}{nN}S^2$$

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

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

$$= \left(\frac{n-1}{nN}\right)S^2 > 0$$

That implies $$\text{Var}_{\text{SRSWR}}(\bar{x}) > \text{Var}_{\text{SRSWOR}}(\bar{x})$$

That means variance of the sample mean is more in SRSWR as compared with its variance in the case of SRSWOR. In other words SRSWOR provides a more efficient estimate of sample mean relative to SRSWR.

2.3.1 Merits and Demerits of Simple Random Sampling

Merits

Simple random sampling has the following merits:

1. In simple random sampling each unit of the population has equal chance to be included in the sample; and
2. Efficiency of the estimates can be found out in simple random sampling because all the estimates are calculated by using the probability theory.

Demerits

Despite merits, simple random sampling has some demerits too viz.

1. An up-to-date frame of population is required in simple random sampling;
2. Some administrative inconvenience arises in simple random sampling if some of the units are spreaded in a wide area. So collecting information from these related units may be problem; and
3. SRS required larger sample size than any other sampling for a fix level of precision.

Example 1: A population have 7 units 1, 2, 3, 4, 5, 6, 7. Write down all possible samples of size 2 (without replacement) which can be drawn from the given population and verify that sample mean is an unbiased estimate of the population mean. Also calculate its sample variance and verify that

$$\text{Var}_{\text{SRSWR}}(\bar{x}) > \text{Var}_{\text{SRSWOR}}(\bar{x})$$

Solution: We have

$$X = 1, 2, 3, 4, 5, 6, 7$$

$$\bar{X} = \frac{1+2+3+4+5+6+7}{7} = 4$$

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

$$=\frac{1}{6}\times(9+4+1+0+1+4+9)$$

$$=\frac{28}{6}=4.666$$

All possible samples of size 2 are as follows:

Sample No.	Sample values	Sample Mean $(\bar{x})$	$\bar{x}-\bar{X}$	$(\bar{x}-\bar{X})^2$
1	(1,2)	1.5	-2.5	6.25
2	(1,3)	2.0	-2.0	4.00
3	(1,4)	2.5	-1.5	2.25
4	(1,5)	3.0	-1.0	1.00
5	(1,6)	3.5	-0.5	0.25
6	(1,7)	4.0	0	0
7	(2,3)	2.5	-1.5	2.25
8	(2,4)	3.0	-1.0	1.00
9	(2,5)	3.5	-0.5	0.25
10	(2,6)	4.0	0	0
11	(2,7)	4.5	+0.5	0.25
12	(3,4)	3.5	-0.5	0.25
13	(3,5)	4.0	0	0
14	(3,6)	4.5	+0.5	0.25
15	(3,7)	5.0	+1.0	1.00
16	(4,5)	4.5	+0.5	0.25
17	(4,6)	5.0	+1.0	1.00
18	(4,7)	5.5	+1.5	2.25
19	(5,6)	5.5	+1.5	2.25
20	(5,7)	6.0	+2.0	4.00
21	(6,7)	6.5	+2.5	6.25
Total		84.0		35.00

From the table, we have

$$\sum \bar{x}_i = 84.0 \text{ and } \sum_{i=1}^{k} (\bar{x}_i - \bar{X})^2 = 35.00$$

$$E(\bar{x}) = \frac{\sum_{i=1}^{{}^{N}C_n} \bar{x}_i}{{}^{N}C_n} = \frac{84}{21} = 4 = \bar{X}$$

$$Var(\bar{x}) = \frac{1}{{}^{N}C_n} \sum_{i=1}^{{}^{N}C_n} (\bar{x}_i - \bar{X})^2 = \frac{1}{21} \times 35.00 = 1.667$$

$$\sigma^2 = \frac{N-1}{N} S^2 = \frac{6}{7} \times 4.667 = 4.0008$$

Verification: In SRSWOR the variance of sample mean is given by

$$\text{Var}(\bar{x}) = \frac{N-n}{Nn}S^2 = \frac{7-2}{7\times 2}4.667 = 1.667$$

In SRSWR the variance of sample mean is given by

$$\text{Var}(\bar{x}) = \frac{\sigma^2}{n} = \frac{4.0008}{2} = 2.0004$$

Hence, $\text{Var}(\bar{x})_{SRSWR} > \text{Var}(\bar{x})_{SRSWOR}$

E1) Draw all possible samples of size 2 from the population {2, 3, 4} and verify that $E(\bar{x}) = \bar{X}$ also find variance.

E2) How many random samples of size 5 can be drawn from a population of size 10 if sampling is done with replacement?

E3) From a population of 50 units, a random sample of size 10 is drawn without replacement. From the sample following result are obtained.

$$\sum_{i=1}^{n} x_i = 48 \ , \ \sum_{i=1}^{n}(x_i - \bar{x})^2 = 36$$

Calculate the sample mean and its variance.

E4) Draw all possible samples of size 2 from the population {8, 12, 16} and verify that

$$E(\bar{x}) = \bar{X}$$

and find variance of estimate of the population mean.

E5) From a population of size N=100, a random sample of size 10 is drawn without replacement. From the sample following results are obtained.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 45$$

Calculate the variance of sample mean.

2.4 SIMPLE RANDOM SAMPLING OF ATTRIBUTES

A qualitative characteristic which cannot be measured numerically is known as an attribute i.e. honesty, intelligence, beauty, etc. In many situations, it is not possible to measure the characteristic under study but possible to classify the population into various classes according to the attributes under study. For example, we can divide a population of a colony into two classes only say literate and illiterate with respect to attribute literacy. Hence the units in the population can be distributed in these two classes accordingly as it possesses or does not possess the given attribute. After taking a sample of size n, we may be interested in estimating the total number of proportion of the defined attribute.

Notations and Terminologies

Let us suppose that a population having N units $X_1, X_2, \ldots, X_N$ is classified into k mutually disjoint and exhaustive classes. Then

π = The proportion of units processing the given attribute in population A/N

π' = The proportion of units not processing the given attribute in population $A'/N = 1 - \pi$

Let us consider SRSWOR sample of size n. From this population if 'a' is the number of units in a sample possessing the given attribute then

p = proportion of sampled units possessing the given attribute = a / n

q = proportion of sampled units not possessing the given attribute = a'/n

Let X_i be the i^{th} unit of the population, where i = 1, 2, …, N.

Then, $X_i = 1$ if i^{th} unit possesses the given attribute

= 0 if it does not possess the given attribute

Similarly, x_i denote i^{th} unit in the sample

Then, $x_i = 1$, if i^{th} sampled unit possesses the given attribute

$x_i = 0$, if i^{th} sampled unit does not possess the given attribute

The $\sum_{i=1}^{N} X_i$ = A, the number of units in the population possessing the given attribute.

and $\sum_{i=1}^{n} x_i$ = a, the number of sampled units possessing the given attributes.

Thus, $$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{A}{N} = \pi$$

and $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{a}{n} = p$$

Similarly,

$$\sum_{i=1}^{N} X_i^2 = A = N\pi$$

and $$\sum_{i=1}^{n} x_i^2 = a = np$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} X_i^2 - N\overline{X}^2\right]$$

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

Similarly,

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2\right]$$

$$= \frac{1}{n-1}\left[np - np^2\right] = \frac{npq}{n-1}$$

Theorem 10: Sample proportion p is an unbiased estimate of population proportion π, i.e. $E(p) = \pi$

Proof: We have

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{a}{n} = p$$

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{A}{N} = \pi$$

We know that in simple random sampling the sample mean provides on unbiased estimate of the population mean

$$E(\bar{x}) = \bar{X}$$

Therefore $E(p) = \pi$

Theorem 11: In SRSWOR, show that $\text{Var}(p) = \frac{N-n}{N-1}.\frac{\pi(1-\pi)}{n}$

Proof: We have, $\text{Var}(p) = \text{Var}(\bar{x})$

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

$$= \frac{N-n}{n.N}.\frac{N.\pi(1-\pi)}{N-1}$$

$$= \frac{N-n}{N-1}.\frac{\pi(1-\pi)}{n}$$

2.5 SAMPLE SIZE FOR SPECIFIC PRECISION

A very first problem faced by a statistician in any sample survey is to determine the sample size so that the population parameters may be estimated with a specified precision. The degree of precision can be determined in terms of

1. The level of significance in the estimate; and
2. The confidence interval within which this estimate lie with respect to given level of significance.

Let us consider the parameter $\bar{X}$ the population mean of the population of size N. We know that $\bar{x}$ sample mean based on n units is unbiased estimate of $\bar{X}$. Let the difference between estimate value $\bar{x}$ and the population mean $\bar{X}$ is d and level of confidence is $(1 - \alpha)$, then the sample size is determined by the equation.

$$P\left[\left|\bar{x}-\bar{X}\right|<d\right]=1-\alpha \quad \ldots (15)$$

or

$$P\left[\left|\bar{x}-\bar{X}\right|\geq d\right]=\alpha \quad \ldots (16)$$

where, α is very small preassinged probability and is nown as the level of significance.·

If n is sufficiently large and we consider SRSWOR, then the statistic

$$Z = \frac{\bar{x}-E(\bar{x})}{SE(\bar{x})} = \frac{\bar{x}-\bar{X}}{\sqrt{Var(\bar{x})}} = \frac{\bar{x}-\bar{X}}{S\sqrt{\frac{1}{n}-\frac{1}{N}}} \quad \ldots (17)$$

where, Z is a standard normal variate.

Accordingly, if we take $\alpha = 0.05$, then we have

$$P\left[\left|Z\right|\geq 1.96\right] = 0.05$$

$$P\left[\left|\frac{\bar{x}-\bar{X}}{S\sqrt{\frac{1}{n}-\frac{1}{N}}}\right|\geq 1.96\right] = 0.05$$

$$P\left[\left|\bar{x}-\bar{X}\right|\geq 1.96\times S\sqrt{\frac{1}{n}-\frac{1}{N}}\right] = 0.05$$

Comparing with equation (16), we get

$$d = 1.96\times S\sqrt{\frac{1}{n}-\frac{1}{N}}$$

$$\frac{d^2}{S^2(1.96)^2} = \frac{1}{n} - \frac{1}{N}$$

$$n = \frac{NS^2(1.96)^2}{Nd^2+S^2(1.96)^2} = \frac{3.84NS^2}{3.84S^2+Nd^2} \quad \ldots (18)$$

This formula gives the sample size in SRSWOR for estimating population mean with confidence level 95 % and margin of error d, provided n is large.

Similarly, if n is small the statistics z follows the student's t distribution with (n-1) degree of freedom is given by

$$t = \frac{\bar{x}-\bar{X}}{S\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)}}$$

If t_α is the critical value of t for (n−1) df and at α level of significance then n is given by the equation

$$P\left[\left|\bar{x}-\bar{X}\right| \geq S\sqrt{\frac{1}{n}-\frac{1}{N}}.t_\alpha\right] = \alpha \qquad \ldots (19)$$

Comparing with equation (16) we get

$$d = S\sqrt{\left(\frac{1}{n}-\frac{1}{N}\right)}.t_\alpha$$

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

$$n = \frac{NS^2\ t_\alpha^2}{Nd^2+S^2t_\alpha^2} = \frac{S^2t_\alpha^2}{d^2+\left(S^2\ t_\alpha^2 \big/ N\right)}$$

2.6 SUMMARY

In this unit, we have discussed:

1. The simple random sampling;
2. The method of SRSWR and SRSWOR;
3. The properties of simple random sampling;
4. Method of finding the variance of the estimate of the sample mean;
5. The simple random sampling for attribute and its properties; and
6. The sample size determination for specific precision.

2.7 SOLUTIONS / ANSWERS

E1) In SRSWOR the number of sample is

$${}^{N}C_n = {}^{3}C_2 = 3$$

The samples with their means are as follows:

Sr. No	Sample	Mean $(\bar{x})$
1	2,3	2.5
2	2,4	3
3	3,4	3.5
		$\sum_{i=1}^{3}\bar{x}_i = 9$

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

$$= \frac{9}{3} = 3$$

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

$$= \frac{1}{3}(2+3+4)$$

$$= \frac{9}{3} = 3$$

Therefore,

$$E(\bar{x}) = \bar{X}$$

Again $$V(\bar{x}) = \frac{N-n}{N.n} S^2$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

$$= \frac{1}{2} \left\{(2-3)^2 + (3-3)^2 + (4-3)^2\right\}$$

$$=1$$

Therefore,

$$Var(\bar{x}) = \frac{3-2}{3\times 2} \times 1$$

$$= \frac{1}{6} = 0.166$$

E2) The first unit can be drawn from 10 units in ${}^{10}C_1 = 10$ ways. Since sampling is done with replacement so the second unit can be drawn in ${}^{10}C_1 = 10$ ways … so on upto the selection of 5^{th} Unit. Thus the total ways are $10.10.10.10.10 = 10^5$ ways.

E3) We have

$$\sum_{i=1}^{n} x_i = 48$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{10} \times 48 = 4.8$$

So

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

$$s^2 = \frac{1}{9} \times 36 = 4$$

which is the estimate value of S^2.

Therefore,

$$\text{Variance}(\bar{x}) = \frac{N-n}{Nn}S^2 = \frac{50-10}{50\times 10}\times 4$$
$$= \frac{16}{50} = 0.32$$

E4) In SRSWOR the number of samples is ${}^{N}C_n = {}^{3}C_2 = 3$ and samples with their means are

S. No.	Sample	Mean $(\bar{x}_i)$
1	(8, 12)	10
2	(8, 16)	12
3	(12, 16)	14
Total		36

$$\sum_{i=1}^{3}\bar{x}_i = 36 \ ,$$

Therefore,

$$E(\bar{x}) = \frac{1}{{}^{N}C_n}\sum_{i=1}^{{}^{N}C_n}\bar{x}_i$$
$$= \frac{1}{3}\left(\sum_{i=1}^{3}\bar{x}_i\right)$$
$$= \frac{1}{3}\times 36 = 12$$

Again $$\bar{X} = \frac{8+12+16}{3} = 12$$

Therefore,

$$E(\bar{x}) = \bar{X}$$

Again estimator of population mean is sample mean and so its variance

$$\text{Var}(\bar{x}) = \left(\frac{1}{n}-\frac{1}{N}\right)S^2$$

where,

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2$$
$$= \frac{1}{3-1}\left[(8-12)^2+(12-12)^2+(16-12)^2\right]$$
$$= \frac{1}{2}\left[16+0+16\right] = 16$$

Therefore,

$$\text{Var}(\bar{x}) = \left(\frac{1}{2}-\frac{1}{3}\right)\times 16$$

$$=\left(\frac{3-2}{6}\right)16 = \frac{16}{6}$$

$$= \frac{8}{3} = 2.66$$

E5) We have,

$$\text{Var}(\bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

and

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

$$= \frac{1}{9} \times 45 = 5$$

Which is the estimate value of S^2.

Therefore,

$$\text{Var}(\bar{x}) = \left(\frac{1}{10} - \frac{1}{100}\right) \times 5$$

$$= \frac{9}{100} \times 5 = 0.45$$

UNIT 3 STRATIFIED RANDOM SAMPLING

Structure

3.1 Introduction
Objectives
3.2 Principles of Stratification
Notations and Terminology
3.3 Properties of Stratified Random Sampling
3.4 Mean and Variance for Proportions
3.5 Allocation of Sample Size
Equal Number of Units from Each Stratum
Proportional Allocation
Neyman's Allocation
Optimum Allocation
3.6 Stratified Sampling versus Simple Random Sampling
Proportional Allocation Versus Simple Random Sampling
Neyman's Allocation Versus Proportional Allocation
Neyman's Allocation Versus Simple Random Sampling
Merits and Demerits of Stratified Random Sampling
3.7 Summary
3.8 Solutions/Answers

3.1 INTRODUCTION

When the units of the population are scattered and not completely homogeneous in nature, then simple random sample does not give proper representation of the population. So if the population is heterogeneous the simple random sampling is not found suitable. In simple random sampling the variance of the sample mean is proportional to the variability of the sampling units in the population. So, in spite of increasing the sample size n or sampling fraction n/N, the only other way of increasing the precision is to device a sampling which will effectively reduce the variability of the sample units, the population heterogeneity. One such method is stratified sampling method.

In stratified sampling the whole population is to be divided in some homogeneous groups or classes with respect to the characteristic under study which are known as strata. That means, we have to do the stratification of the population. Stratification means division into layers. The auxiliary information related to the character under study may be used to divide the population into various groups or strata in such a way that units within each stratum are as homogeneous as possible and the strata are as widely different as possible.

Thus, all strata would comprise the population. Then from each stratum sample would be drawn and lastly all samples would be combined to get the ultimate sample. For example, let us consider that population consists of N units and these are distributed in a heterogeneous structure. Now first of all

we divide the population into 'k' non overlapping strata of sizes N_1, N_2, N_3, ..., N_k such that each stratum becomes homogeneous. Evidently $N = N_1 + N_2 + N_3 + ... + N_k$. Then from first stratum a sample of size n_1 would be drawn by simple random sampling method. Similarly, from the second stratum a sample of n_2 units would be drawn and so on, up to k^{th} stratum. Now all these k samples would be combined to get the ultimate sample. So, the ultimate size of sample would be $n = n_1 + n_2 + n_3 + ... + n_k$. This method of sampling is known as Stratified random sampling because here stratification is done first to make population homogeneous and then samples are drawn randomly by simple random sampling from each stratum.

The principles of stratification are explained in Section 3.2. The properties of stratified random sampling are described in Section 3.3, whereas Section 3.4 provides the derivation of the mean and variance of proportions in stratified random sampling. The allocation of sample size with the help of different techniques is described in Section 3.5. The comparative study between stratified random sampling and simple random sampling is given in Section 3.6.

Objectives

After studying this unit, you would be able to

- define the stratified random sampling;
- explain the principles of stratification;
- describe the properties of stratified random sampling;
- derive the mean and variance of proportions in stratified random sampling;
- describe the allocation of sample size with the help of different techniques; and
- calculate the estimate of population mean and variance of sample mean.

3.2 PRINCIPLES OF STRATIFICATION

The principles to be kept in mind while stratifying a population are given below:

1. The strata should not be overlapping and should together comprise the whole population.

2. The strata should be homogeneous within themselves and heterogeneous between themselves with respect to characteristic under study.

3. If a investigator is facing difficulties in stratifying a population with respect to the characteristic under study, then he/she has to consider the administrative convenience as the basis for stratification.

4. If the limit of precision is given for certain sub-population then it should be treated as stratum.

3.2.1 Notations and Terminology

N = Population size

n = Sample size

k = Number of strata

N_i = Size of i^{th} stratum

Then $N = \sum_{i=1}^{k} N_i$

n_i = Size of sample drawn from i^{th} stratum

Then $n = \sum_{i=1}^{k} n_i$

X_{ij} = Value of character under study for j^{th} unit of i^{th} stratum

$\overline{X}_i$ = Population mean of i^{th} stratum $= \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$

$\overline{X}$ = Population Mean $= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} X_{ij}$

$= \frac{1}{N}\sum_{i=1}^{k} N_i \overline{X}_i = \sum_{i=1}^{k} W_i \overline{X}_i$

where, $W_i = \frac{N_i}{N}$ is called the weight of i^{th} stratum

S_i^2 = Population mean square of the i^{th} stratum

$= \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(X_{ij} - \overline{X}_i\right)^2$, $\left(j = 1, 2, ..., N_i \ \& \ i = 1, 2, ..., \ k\right)$

x_{ij} = Value of j^{th} sample unit taken from i^{th} stratum

$\overline{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ = Mean of sample units selected from i^{th} stratum

$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}_i\right)^2$, $\left(i = 1, 2, ..., k\right)$

= Sample mean square of the i^{th} stratum

Let us consider the following sample means to estimate the populations mean of i^{th} stratum $\overline{X}$ which are:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{x}_i$$

and $$\overline{x}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \overline{x}_i = \sum_{i=1}^{k} W_i \overline{x}_i$$

where, $\overline{x}_{st}$ is the weighted mean of the strata sample means, weights being equal to strata sizes. These two will be identical if $n_i \propto N_i$

3.3 PROPERTIES OF STRATIFIED RANDOM SAMPLING

Theorem 1: $\bar{x}_{st}$ is an unbiased estimate of the population mean $\bar{X}$ i.e.

$$E(\bar{x}_{st}) = \bar{X}$$

Proof: We have

$$\bar{x}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{x}_i$$

Therefore,

$$E(\bar{x}_{st}) = E\left[\frac{1}{N}\sum_{i=1}^{k} N_i \bar{x}_i\right]$$

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

Since the sample units selected from each of stratum are simple random sample, then we have

$$E(\bar{x}_i) = \bar{X}_i$$

Therefore,

$$E(\bar{x}_{st}) = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{X}_i$$

$$= \frac{1}{N}\sum_{i=1}^{k} N_i \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$$

$$= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} X_{ij} = \bar{X}$$

Hence proved

Theorem 2: Prove that

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

Proof: We have

$$Var(\bar{x}_{st}) = Var\left(\sum_{i=1}^{k} W_i \bar{x}_i\right)$$

$$= \sum_{i=1}^{k} W_i^2 \, Var(\bar{x}_i)$$

The covariance term vanish since the samples from different strata are independent and the sample units in each stratum are the simple random sample without replacement, we have

$$\text{Var}\left(\bar{x}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

or $$\text{Var}\left(\bar{x}_i\right) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^{\,2}$$

Therefore,

$$\text{Var}\left(\bar{x}_{st}\right) = \sum_{i=1}^{k} W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2$$

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

From the above result the variance depends on S_i^2 the heterogeneity within the strata. Thus, if S_i^2 are small i.e. strata are homogeneous then stratified sampling schemes provides estimates with greater precision.

Theorem 3: If S_i^2 is not known then prove that estimate of the variance of the sample mean of the stratified random sample is given by

$$E\left[\text{Var}\left(\bar{x}_{st}\right)\right] = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2$$

Proof: In general S_i^2 are not known. A simple random sample is drawn from each stratum. If we assume a individual stratum as a population then the sample, drawn from it, would be a simple random sample. If the sample is drawn from i^{th} stratum, the sample mean square s_i^2 would be an estimate of population mean square S_i^2

$$\text{i.e.} \quad E\left(s_i^2\right) = S_i^2 \qquad i = 1, 2, ..., k \qquad \ldots (1)$$

Accordingly an unbiased estimate of the variance is given by

$$\text{Var}\left(\bar{x}_{st}\right) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 \, s_i^2$$

Therefore,

$$E\left[\text{Var}\left(\bar{x}_{st}\right)\right] = E\left[\sum_{i=1}^{n} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 \, s_i^2\right]$$

$$= \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 \, E\left(s_i^2\right)$$

Substituting from equation (1), we get

$$E\left[\text{Var}\left(\bar{x}_{st}\right)\right] = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 \, S_i^2$$

3.4 MEAN AND VARIANCE FOR PROPORTIONS

As in simple random sampling, we can divide a population into two classes with respect to a attribute. Hence the units in the population are classified in these two classes accordingly as it possesses or does not possess the given attribute. After taking a sample of size n, we may be interested in estimating the population proportion of the defined attribute.

If a unit possesses the attribute, it receives the code value 1 and if an unit does not possesses the attribute, it receives the value 0. Let the number of units belonging to A in the i^{th} stratum of size N_i be M_i and if the sample of size n_i taken from i^{th} stratum, the number of units belonging to A be m_i. Denoting the proportion of units belonging to A in the population, in the i^{th}stratum and sample from the i^{th} stratum by π, π_i and p_i respectively, various formula for mean and variance are as follows:

$$\pi_i = \frac{M_i}{N_i} \text{ and } p_i = \frac{m_i}{n_i}$$

and

$$\pi = \sum_{i=1}^{k} \frac{N_i}{N} \pi_i = \sum_{i=1}^{k} W_i \pi_i \qquad \text{for } i = 1, 2, \ldots, k$$

The estimated proportion p_{st} under stratified sampling for the units belonging to A is

$$p_{st} = \sum_{i=1}^{k} W_i p_i$$

$$\text{Mean}(p_{st}) = E\left(\sum_{i=1}^{k} W_i p_i \right) = \sum_{i=1}^{k} W_i E(p_i)$$

since we draw SRS from each stratum so by Theorem 10 of Section 2.4 of Unit 2 we have

$$E(p_i) = \pi_i$$

By putting this value in above formula, we have

$$\text{Mean}(p_{st}) = \sum_{i=1}^{k} W_i \pi_i = \pi$$

Hence sample proportion under stratified sampling is unbiased estimate for population proportion.

Now variance of p_{st} is given by

$$\text{Var}(p_{st}) = \text{Var}\left(\sum_{i=1}^{k} W_i p_i \right) = \frac{1}{N^2} \sum_{i=1}^{k} N_i^2 V(p_i)$$

since we draw SRS from each stratum so by Theorem 11 of Section 2.4 of Unit 2 we have

$$\mathrm{Var}(p_i) = \frac{N_i - n_i}{N_i - 1} \cdot \frac{\pi_i(1-\pi_i)}{n_i}$$

By putting this value in above formula, we have

$$\mathrm{Var}(p_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} \frac{N_i^{\,2}(N_i - n_i)\pi_i(1-\pi_i)}{(N_i - 1)n_i}$$

If N_i is large enough, consider $1/N_i$ as negligible and $N_i - 1 \sim N_i$, formula for $\mathrm{Var}(p_{st})$ reduces to

$$\mathrm{Var}(p_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} N_i (N_i - n_i) \frac{\pi_i(1-\pi_i)}{n_i}$$

and if n_i / N_i is negligible, therefore

$$\mathrm{Var}(p_{st}) = \sum_{i=1}^{k} W_i^2 \frac{\pi_i(1-\pi_i)}{n_i}$$

An unbiased estimate of $\mathrm{Var}(p_{st})$ is given by

$$E\left[\mathrm{Var}(p_{st})\right] = \frac{1}{N^2} \sum_{i=1}^{k} N_i \frac{N_i - n_i}{(n_i - 1)} p_i q_i$$

where, $q_i = 1 - p_i$

3.5 ALLOCATION OF SAMPLE SIZE

In stratified sampling, the allocation of the sample to different strata is done by considering the following factors:

1. The total number of units in the stratum, i.e. stratum size;
2. The variability within the stratum; and
3. The cost in taking observations per sampling unit in the stratum.

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

Types of Allocation of Sample Size

It is evident from the formula for variance of $\bar{x}_{st}$ that it depends on n_i the number of units selected at random from i^{th} stratum. Hence, the problem arises, what optimum value of $n_i\ (i = 1, 2, ..., k)$ can be chosen out of n, so that, the variance is as small as possible. Four types of allocations are considered here:

3.5.1 Equal Number of Units from Each Stratum

This is a situation of considerable practical interest for reasons of administrative convenience. In this allocation method, the total sample size n is divided equally among all the strata i.e. if the population is divided in k strata then the size of sample for each stratum would be

$$n_i = \frac{n}{k} \qquad \text{for all } i = 1, 2, \ldots, k$$

3.5.2 Proportional Allocation

This allocation was originally proposed by Bowley in 1926. This procedure of allocation is very common in practice because of its simplicity. As its name indicates, proportional allocation means that we select a small sample from a small stratum and a large sample from a large stratum. The sample size in each stratum is fixed in such a way that for all the strata, the ratio n_i/N_i is equal to n/N i.e.

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

$$\text{or} \quad n_i = \frac{n}{N} N_i$$

$$n_i = n W_i$$

$$\text{or} \quad n_i \propto N_i \qquad \ldots (2)$$

In other words, the allocation of a sample of size n to different strata is to be done in proportional to their sizes. We have variance of $\bar{x}_{st}$

$$\text{Var}(\bar{x}_{st}) = \sum_{i=1}^{k} \left(1 - \frac{n_i}{N_i}\right) \frac{W_i^2}{n_i} S_i^2$$

Thus, substituting the values of n_i and n_i/N_i as $n W_i$ and n/N respectively in variance formula then we get variance under proportional allocation

$$\text{Var}(\bar{x}_{st})_{PROP} = \left(1 - \frac{n}{N}\right) \sum_{i=1}^{k} \frac{W_i S_i^2}{n}$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i S_i^2 \qquad \ldots (3)$$

In case, the sampling fraction (n/ N) is negligible

$$\text{Var}(\bar{x}_{st})_{PROP} = \sum_{i=1}^{k} \frac{W_i S_i^2}{n} \qquad \ldots (4)$$

3.5.3 Neyman's Allocation

This allocation of the total sample size n to the different stratum is called minimum variance allocation and is due to Neyman (1934). This result was first discovered by Tchuprow (1923) but remained unknown until it was rediscovered independently by Neyman. This allocation of samples among different strata is based on a joint consideration of the stratum size and the stratum variance. In this allocation, it is assumed that the sampling cost per unit among different strata is same and the size of the sample is fixed. Sample sizes are allocated by

$$n_i = n\frac{W_i S_i}{\sum_{i=1}^{k} W_i S_i} = n\frac{N_i S_i}{\sum_{i=1}^{k} N_i S_i} \quad \ldots (5)$$

A formula for the minimum variance with fixed n is obtained by substituting the value of n_i in variance formula, then we get

$$Var(\bar{x}_{st})_{NEY} = \frac{\left(\sum_{i=1}^{k} W_i S_i\right)^2}{n} - \frac{\sum_{i=1}^{k} W_i S_i^2}{N} \quad \ldots (6)$$

3.5.4 Optimum Allocation

The variance of estimated mean depends on n_i which can arbitrarily be fixed. One more factor, which is none the less important, also influences the variance of estimated mean. The allocation problem is two fold:

1. We attain maximum precision for the fixed cost of the survey; and
2. We attain the desired degree of precision for the minimum cost.

Thus, the allocation of the sample size in various strata, in accordance with these two objectives, is known as optimum allocation.

In any stratum the cost of survey per sampling unit cannot be the same. That is, in one stratum the cost of transportation may be different from the other. Hence, it would not be wrong to allocate the cost of the survey in each stratum differently.

Let c_i be the cost per unit of survey in the i^{th}stratum from which a sample of size n_i is stipulated. Also suppose c_0 as the over head fixed cost of the survey. In this way the total cost C of the survey comes out to be

$$C = c_0 + \sum_{i=1}^{k} c_i n_i \quad \ldots (7)$$

c_0 and c_i are beyond our control. Hence we will determine the optimum value of n_i which minimizes the variance of stratified sample mean.
To determine the optimum value of n_i, we consider the function

$$\begin{aligned}\varphi &= Var(\bar{x}_{st}) + \lambda C \\ &= \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) W_i^2 S_i^2 + \lambda\left(c_0 + \sum_{i=1}^{k} c_i n_i\right)\end{aligned} \quad \ldots (8)$$

where λ is constant and known as Lagrange's multiplier.

Using the method of Lagrange's multiplier we select n_i and the constant λ to minimize φ. Differentiating φ with respect to n_i, we have

$$-\frac{W_i^2 S_i^2}{n_i^2} + \lambda c_i = 0 \qquad \ldots (9)$$

or

$$n_i = \frac{W_i\ S_i}{\sqrt{\lambda\, c_i}} \qquad \ldots (10)$$

Since λ is an unknown quantity, it has to be determined interms of known values. So, we take the sum over all in equation (10) and thus obtain

$$\sum_{i=1}^{k} n_i = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{k} \frac{W_i S_i}{\sqrt{c_i}}$$

or $$n = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{k} \frac{W_i S_i}{\sqrt{c_i}}$$

$$\Rightarrow \sqrt{\lambda} = \frac{1}{n} \sum_{i=1}^{k} \frac{W_i S_i}{\sqrt{c_i}}$$

Substituting the value of $\sqrt{\lambda}$ in equation (10), we get the value of n_i

$$n_i = n \frac{\left(W_i S_i / \sqrt{c_i}\right)}{\sum_{i=1}^{k} \left(W_i S_i / \sqrt{c_i}\right)} = n \frac{\left(N_i S_i / \sqrt{c_i}\right)}{\sum_{i=1}^{k} \left(N_i S_i / \sqrt{c_i}\right)} \qquad \ldots (11)$$

Thus, the relation (11) leads to the following important conclusions that we have to take a larger sample in a given stratum if

1. The stratum size N_i is larger;
2. The stratum has larger variability (S_i); and
3. The cost per unit is lower in the stratum.

3.6 STRATIFIED SAMPLING VERSUS SIMPLE RANDOM SAMPLING

Now, we shall make a comparative study of simple random sampling without replacement and stratified random sampling under different kinds of allocations i.e. Proportional allocation and Neyman's allocation.

3.6.1 Proportional Allocation versus Simple Random Sampling

The variance of the estimate of stratified sample mean with proportional allocation and variance of the sample mean of simple random sampling is given respectively by

$$\text{Var}\ (\bar{x}_{st})_{PROP} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i S_i^2 \qquad \ldots (12)$$

where, $S_i^2 = \frac{1}{(N_i - 1)} \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2$

and $$Var(\bar{x})_{SRSWOR} = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \quad \ldots (13)$$

where, $S^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{N_i} (X_{ij} - \bar{X})^2$

In order to comparing (12) and (13) we shall first express S^2 in terms of S^2_i we have

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i + \bar{X}_i - \bar{X})^2$$

$$(N-1)S^2 = \sum_{i=1}^{k} \left[\sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2 \right] + \sum_{i=1}^{k} \sum_{j=1}^{N_i} (\bar{X}_i - \bar{X})^2 + 2 \sum_{i=1}^{k} \left[(\bar{X}_i - \bar{X}) \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i) \right]$$

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

The product term vanishes since

$$\sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i) = 0$$

being the sum of square of deviation from the stratum mean. If we assume that N_i and consequently N are sufficiently large so that we can put $N_i - 1 = N_i$ and $N - 1 = N$, then we get

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

$$S^2 = \sum_{i=1}^{k} W_i S_i^2 + \sum_{i=1}^{k} W_i (\bar{X}_i - \bar{X})^2 \quad \ldots (14)$$

Substituting in equation (13), we get

$$Var(\bar{x})_{SRSWOR} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i S_i^2 + \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i (\bar{X}_i - \bar{X})^2$$

$$Var(\bar{x})_{SRSWOR} = Var(\bar{x}_{st})_{PROP} + \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} W_i (\bar{X}_i - \bar{X})^2$$

$$Var(\bar{x})_{SRSWOR} \geq Var(\bar{x}_{st})_{PROP} \quad \ldots (15)$$

3.6.2 Neyman's Allocation versus Proportional Allocation

Considering the variances of estimated sample mean in stratified random sampling with proportional allocation and Neyman's allocation respectively we have

$$\text{Var}\left(\bar{x}_{st}\right)_{PROP} = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i S^2 \quad \ldots (16)$$

and

$$\text{Var}\left(\bar{x}_{st}\right)_{NEY} = \frac{1}{n}\left(\sum_{i=1}^{k} W_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_i S_i^2 \quad \ldots (17)$$

By subtracting equation (17) from equation (16) we get

$$\begin{aligned}\text{Var}\left(\bar{x}_{st}\right)_{PROP} - \text{Var}\left(\bar{x}_{st}\right)_{NEY} &= \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i S_i^2 \\ &\quad - \left\{\frac{1}{n}\left(\sum_{i=1}^{k} W_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} W_i S_i^2\right\} \\ &= \frac{1}{n}\left[\sum_{i=1}^{k} W_i S_i^2 - \left(\sum_{i=1}^{k} W_i S_i\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{k} W_i (S_i - \bar{S})^2 \end{aligned} \quad \ldots (18)$$

where, $\bar{S} = \sum_{i=1}^{k} W_i S_i = \frac{1}{N}\sum_{i=1}^{k} N_i S_i$ is the weighted mean of the stratum sizes N_i

Hence from equation (18) we can say

$$\text{Var}\left(\bar{x}_{st}\right)_{PROP} \geq \text{Var}\left(\bar{x}_{st}\right)_{NEY}$$

because, R.H.S. of equation (18) is non-negative.

3.6.3 Neyman's Allocation versus Simple Random Sampling

From the relationship between the proportional allocation and simple random sampling and the relation between proportional and Neyman allocation we have

$$\text{Var}\left(\bar{x}\right)_{SRSWOR} = \text{Var}\left(\bar{x}_{st}\right)_{PROP} + \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i \left(\bar{X}_i - \bar{X}\right)^2 \quad \ldots (19)$$

and $$\text{Var}\left(\bar{x}_{st}\right)_{PROP} = \text{Var}\left(\bar{x}_{st}\right)_{NEY} + \frac{1}{n}\sum_{i=1}^{k} W_i \left(S_i - \bar{S}\right)^2 \quad \ldots (20)$$

By substituting the value of the variance under proportional allocation in equation (19) from equation (20), we have

$$\begin{aligned}\text{Var}\left(\bar{x}\right)_{SRSWOR} &= \text{Var}\left(\bar{x}_{st}\right)_{NEY} + \frac{1}{n}\sum_{i=1}^{k} W_i \left(S_i - \bar{S}\right)^2 \\ &\quad + \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} W_i \left(\bar{X}_i - \bar{X}\right)^2 \end{aligned} \quad \ldots (21)$$

That means

$$\text{Var}\left(\bar{x}\right)_{SRSWOR} \geq \text{Var}\left(\bar{x}_{st}\right)_{NEY}$$

because both the terms in R.H.S. of equation (21) are positive. From the results of the relations of variance of simple random sample mean and the

variance of stratified sample means with proportional and Neyman allocations, we can reach on the conclusion that

$$\text{Var}\left(\bar{x}\right)_{\text{SRSWOR}} \geq \text{Var}\left(\bar{x}_{st}\right)_{\text{PROP}} \geq \text{Var}\left(\bar{x}_{st}\right)_{\text{NEY}}$$

3.6.4 Merits and Demerits of Stratified Random Sampling

Merits

1. **More Representative**
   Stratified random sampling ensures any desired representation in the sample of the various strata in the population. It overruled the probability of any essential group of the population being completely excluded in the sample.

2. **Greater Accuracy**
   Stratified random sampling provides estimate of parameters with increased precision in comparison to simple random sampling. Stratified random sampling also enables us to obtain the results of known precision for each of the stratum.

3. **Administrative Convenience**
   The stratified random samples would be more concentrated geographically in comparison to simple random samples. Therefore, this method needs less time and money involved in interviewing the supervision of the field work can be done with greater case and convenience.

Demerits

However, stratified random sampling has some demerits too, which are:

1. **May Contain Error due to Subjectiveness**
   In stratified random sampling the main objective is to stratify the population in homogeneous strata. But stratification is a subjective issue and so it may contain error.

2. **Lower Efficiency**
   If the sizes of samples from different stratum are not properly determined then stratified random sampling may yield a larger variance that means lower efficiency.

Example 1: A sample of 60 persons is to be drawn from a population consisting of 600 belonging to two villages A and B. The means and standard deviations of their marks are given below:

Villages	**Strata Sizes** (N_i)	**Means** $(\bar{X}_i)$	**Standard Deviation** (σ_i)
Village A	400	60	20
Village B	200	120	80

Draw a sample using proportional allocation techniques.

Solution: If we regard the villages A and B as representing two different strata then the problem is to draw a stratified random sample of size 30 using technique of proportional allocation. In proportional allocation, we have

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

Therefore,

$$n_1 = \frac{60}{600} \times 400 \ = 40$$

$$n_2 = \frac{60}{600} \times 200 \ = 20$$

Thus, the required sample sizes for the villages A and B are 40 and 20 respectively.

E1) A sample of 100 employees is to be drawn from a population of collages A and B. The population means and population mean squares of their monthly wages are given below:

Village	N_i	$\overline{X}_i$	S_i^2
Collage A	300	25	25
Collage B	200	50	100

Draw the samples using proportional and Neyman allocation technique and compare.

E2) Obtain the sample mean and estimate of the population mean for the given information in Example 1 discussed above.

3.7 SUMMARY

In this unit, we have discussed:

1. The definition and procedure of stratified random sampling;
2. The principles of stratification;
3. The properties of stratified random sampling;
4. The mean and variance of proportions in stratified random sampling;
5. The allocation of sample size with the help of different techniques; and
6. Calculation of the estimate of population mean and variance of sample mean.

3.8 SOLUTIONS /ANSWERS

E1) If we regard the collages A and B representing two different strata then the problem is to draw as stratified sample of 100 employees using technique of proportional allocation and Neyman's allocation.

In proportional allocation we have

$$n_i = \frac{n}{N} N_i = \frac{100}{500} \times N_i$$

$$n_1 = \frac{100}{500} \times 300 = 60$$

$$n_2 = \frac{100}{500} \times 200 = 40$$

In Neyman's allocation, we have

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

$$\sum_{i=1}^{2} N_i S_i = 300 \times 5 + 200 \times 10$$

$$= 1500 + 2000 = 3500$$

$$n_i = 100 \times \frac{N_i S_i}{3500}$$

$$n_1 = 100 \times \frac{300 \times 5}{3500} = 42.85 \cong 43$$

$$n_2 = 100 \times \frac{200 \times 5}{3500} = 57.14 \cong 57$$

Therefore, the samples regarding the colleges A and B for both allocations are obtained as:

	Proportional	**Neyman**
Collage A	60	43
Collage B	40	57
Total	100	100

E2) We have the following data:

Village	N_i	$\bar{X}_i$	σ_i	$S_i^2 = \frac{N}{N-1}\sigma_i^2$	$N_i S_i^2$	$N_i \sigma_i^2$	$\bar{X}_i^2$	$N_i \bar{X}_i^2$
A	400	60	20	400.67	160267.11	160000	3600	3440000
B	200	120	80	6432.16	1286432.16	1280000	14400	880000
Totat	600				1446699.27	1440000	180000	4320000

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{k} N_i \bar{X}_i$$

$$= \frac{1}{600}\left[400\times 60 + 200\times 120\right]$$

$$= \frac{1}{600}\left[24000 + 24000\right] = \frac{48000}{600} = 80$$

$$\text{Var}\left(\bar{x}\ \right)_{\text{PROP}} = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\sum_{i=1}^{k} N_i S_i^2}{N}$$

$$= \left(\frac{600-60}{60\times 600}\right)\times\frac{1446699.27}{600} = \frac{1446699.27}{40000}$$

$$= 36.1675$$

$$\text{Var}\left(\bar{x}\right)_{\text{SRSWOR}} = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

where,

$$S^2 = \frac{1}{(N-1)}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)^2$$

$$= \frac{1}{(N-1)}\left[\sum_{i=1}^{k} N_i\sigma_i^2 + \sum_{i=1}^{k} N_i\left(\bar{X}_i - \bar{X}\right)^2\right]$$

$$= \frac{1}{(N-1)}\left[\sum_{i=1}^{k} N_i\sigma_i^2 + \sum_{i=1}^{k} N_i\bar{X}_i^2 - N\bar{X}^2\right]$$

$$= \frac{1}{599}\left(1440000 + 4320000 - 600\times 80\times 80\right)$$

$$= \frac{1}{599}\left[5760000 - 3840000\right] = \frac{1920000}{599}$$

$$\text{Var}\left(\bar{x}\right)_{\text{SRSWOR}} = \frac{600-60}{600\times 60}\times\frac{1920000}{599}$$

Then the conclusion is

$$\text{Var}\left(\bar{x}_{st}\right)_{\text{PROP}} = 36.1675$$

$$\text{Var}\left(\bar{x}\right)_{\text{SRSWOR}} = 48.08$$

Therefore, precision of $\bar{x}_{st}$ can be obtained by

$$\text{Gain in precision} = \frac{\text{Var}\left(\bar{x}\right)_{\text{SRS}} - \text{Var}\left(\bar{x}_{st}\right)_{\text{PROP}}}{\text{Var}\left(\bar{x}_{st}\right)_{\text{PROP}}}\times 100$$

$$= \frac{48.04 - 36.1675}{36.1675}\times 100$$

$$= 32.8\ \%$$

UNIT 4 SOME OTHER SAMPLING SCHEMES

Structure

4.1 Introduction

Objectives

4.2 Introduction to Systematic Sampling

4.3 Methods of Systematic Sampling

Linear Systematic Sampling

Circular Systematic Sampling

Advantages and Disadvantages of Systematic Sampling

4.4 Properties of Systematic Sampling

Comparision between Simple Random, Stratified and Systematic Sampling

4.5 Introduction to Cluster Sampling

Significance of Systematic Sampling

Formation of Cluster

4.6 Properties of Cluster Sampling

4.7 Introduction to Two-stage Sampling

4.8 Properties of Two-stage Sampling

4.9 Summary

4.10 Solutions / Answers

4.1 INTRODUCTION

When one has to make an inference about a large lot and it is not practically possible to examine each individual unit.Then a few units of the lot are examined and on the basis of the information of those units, one makes decisions about whole lot. In previous unit, we have discussed about stratified random sampling. As we stated in Unit 3, a stratified random sample is selected randomly from the groups of population units named strata, where the strata are formed of the homogeneous units. But, if the formation of the strata is not done in proper manner, then the results would be biased. So, to overcome this drawback, it is suggested to adopt another sampling method known as Systematic sampling. The systematic random sampling is one of among the mixed sampling schemes, which is partly probabilistic and partly non-probabilistic.

A brief introduction of systematic sampling is given in Section 4.2, whereas the methods of selection of systematic random samples are explained with examples in Section 4.3. Properties of the systematic random sampling in terms of mean and variance are derived in Section 4.4. In Section 4.5 the notations and the methods of cluster sampling are explained whereas the properties of the cluster sampling are described in Section 4.6. Similarly, the brief introduction about the two stage sampling is given in Section 4.7 and the properties of the two-stage sampling are discussed in Section 4.8.

Objectives

After studying this unit, you would be able to

- define the systematic random sampling;
- draw a systematic random sample and estimate the population mean;
- obtain the variance of the estimate of the population mean;
- define the cluster sampling and obtain the estimate of population mean; and
- define and explain the two-stage sampling scheme.

4.2 INTRODUCTION TO SYSETMATIC SAMPLING

In previous units, we have discussed those sampling techniques where units were selected randomly. In this unit, we shall discuss a sampling technique which has a nice feature of selecting the whole sample with just one random start. Systematic random sampling is commonly employed if the complete and up-to-date list of population units is available. In systematic random sampling only the first unit is selected with the help of random method and the rest being automatically selected according to some predetermined pattern. The systematic random sampling is a kind of mixed sampling, which is partly probabilistic and partly non-probabilistic. This is random since the first unit of the sample is selected at random and non-random or purposive since the rest of units in the sample are selected by predesigned pattern.

4.3 METHODS OF SYSTEMATIC SAMPLING

Generally, in day to day life, we have to obtain the information from cards or register which are full of information arranged in serial order. For example, the books in library, a telephone directory, etc. In such case systematic sampling often works better than simple random or stratified sampling. There are two ways of selection of a systematic random sample and these are (a) Linear systematic sampling and (b) Circular systematic sampling.

4.3.1 Linear Systematic Sampling

In order to draw a systematic random sample of size n from a population of N units, Let us suppose that N sampling units are serially numbered from 1 to N in some order.

Let N = nk, where n is the sample size and k is an integer. Therefore, $k = N/n$. In other words, in order to draw a sample of size n from N we divide the total number of units into n equal parts. Suppose each part consists of k units. From the serially arranged 1 to k units, draw a unit with random method. Let the selected unit is i^{th} unit where $i \leq k$. The selected unit would be the first sample unit. Then select every k^{th} unit after the i^{th} unit in order to select rest of $(n-1)$ units. Thus, the systematic sample of size n will consists of i^{th}, $(i+k)^{th}$, $(i+2k)^{th}$, ..., $(i+(n-1)k)^{th}$ unit. The random sample unit i.e. the i^{th} unit is called the random start. For example, suppose there are 100 units in a population serially numbered 1 to 100 units. We will divide the whole

population into 10 equal parts of 10 units each if we have to draw a sample of size10:

$$\frac{N}{n}=\frac{100}{10}=10=k$$

Then we draw a unit randomly from 1 to 10 units and let the selected number is 7 i.e. i = 7. Then we select rest of 9 units in a systematic way i.e.

$(i+k)^{th}$, $(i+2k)^{th}$, $(i+3k)^{th}$, ..., $(i+9k)^{th}$ unit from the serially ordered 100 units. Then the systematic sample consists of 7^{th}, 17^{th}, 27^{th}, 37^{th}, 47^{th}, 57^{th}, 67^{th}, 77^{th}, 87^{th} and 97^{th} units.

4.3.2 Circular Systematic Sampling

Suppose a population consists of N units and from this we have to select a systematic sample of n units. Also, assume that N is a multiple of n i.e. N = nk. The procedure is to select a random number, let it be i such that $1 \leq i \leq k$ and then we have to select i^{th} unit from first k unit and then every k^{th} unit i.e. $(i + k)^{th}$, $(i+2k)^{th}$,…,$(i + (n-1) k)^{th}$ positional units. This sampling technique is known as linear systematic sampling. But in general N does not be always a multiple of n. For example N = 17 and n = 4, then $k=\frac{N}{n}=\frac{17}{4}$ = 4.25 has to be taken as 5. Now we select a random number between 1 and 5 and suppose it is 4. Then the remaining three units to be selected are at positions 9, 14, 19. There is no unit in the population at serial number 19. Hence in this situation we can select a sample of 3 only instead of 4. In this situation, the circular systematic sampling is used.

The circular systematic sampling is used when size of the population N is not a multiple of sample n. In this situation we take N/n as k by rounding off N/n to the nearest integer. With regard to selection of a systematic random sample from N units, we have to select random number from 1 to N. Let this number is i. Now we select every $(i+jk-N)^{th}$ unit, when (i + jk) >N putting j = 1, 2, 3, … till n unit are selected. By using the circular systematic sampling we always get a sample of size n. For the same example discussed above with N = 17, n = 4 and k = 5, let the randomly selected number from 1 to 17 is 8 and latter the 13^{th}, 1^{th} and 6^{th} units are selected. When N = nk, the linear and circular systematic sampling plans become identical.

4.3.3 Advantages and Disadvantages of Systematic Sampling

Advantages

Systematic sampling has some advantages over other sampling schemes which are given as:

1. The systematic sampling is very simple and is not very expensive;
2. The systematic sample is uniformly distributed over the whole population and therefore, all sections of the population are represented in the sample; and
3. The managerial control of field work provides an advantage over other sampling methods.

Disadvantages

Systematic sampling has some disadvantages also along with the advantages which are:

1. The main disadvantage of systematic sampling is that samples are not generally random samples;
2. The sample size is different from that required if N is not a multiple of n;
3. In systematic random sampling the sample mean would not be an unbiased estimate of population mean if N is not a multiple of n;
4. We cannot obtain an unbiased estimate of the variance of the estimate of the population mean since it does not provide sampling error; and
5. It may provide highly biased estimate if the sampling frame has a periodic feature.

4.4 PROPERTIES OF SYSTEMATIC SAMPLING

Let x_{ij} denote the j^{th} member of the i^{th} systematic sample (where, i =1, 2, … k; j=1, 2, … n). $\bar{x}_i$ may be denoted as mean of the i^{th} sample, i.e.

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \qquad (i = 1, 2, ..., k)$$

and $$\bar{x}_{sys} = \frac{1}{k}\sum_{i=1}^{k} \bar{x}_i$$

Then $\bar{X}_i$, $\bar{X}_{..}$ and S^2 may be denoted as population mean and population mean square as follows:

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij} \qquad (i = 1, 2, ..., k)$$

$$\bar{X}_{..} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij} = \frac{1}{k}\sum_{i=1}^{k} \bar{X}_i$$

and $$S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_{..}\right)^2 = \frac{1}{(nk-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}\right)^2$$

Theorem 1: In systematic sampling with interval k sample mean $\bar{x}_{sys}$ is an unbiased estimator of population mean $\bar{X}_{..}$ and its variance is given by

$$Var\left(\bar{x}_{sys}\right) = \frac{N-1}{N}S^2 - \frac{(n-1)k}{N}S_{sys}^2$$

where, $$S_{sys}^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2$$

is the mean square among the units which lie within the same systematic sample.

Proof: We have

$$E(\bar{x}_{sys}) = E\left(\frac{1}{k}\sum_{i=1}^{k}\bar{x}_i\right)$$

$$= E\left[\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{j=1}^{n}x_{ij}\right]$$

$$= \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{j=1}^{n}E(x_{ij})$$

$$= \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}X_{ij}$$

$$= \frac{1}{k}\sum_{i=1}^{k}\bar{X}_i = \bar{X}_{..}$$

Now, we have

$$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_{..})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i+\bar{X}_i-\bar{X}_{..})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{X}_i-\bar{X}_{..})^2$$

Covariance term vanishes, since

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)\,(\bar{X}_i-\bar{X}_{..}) = \sum_{i=1}^{k}\left[(\bar{X}_i-\bar{X}_{..})\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)\right] = 0$$

Therefore,

$$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_{..})^2 + n\sum_{i=1}^{k}(\bar{X}_i-\bar{X}_{..})^2$$

$$(N-1)S^2 = k(n-1)S_{sys}^2 + nk\,Var(\bar{x}_{sys})$$

$$Var(\bar{x}_{sys}) = \frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S_{sys}^2$$

4.4.1 Comparison between Simple Ranodm, Stratified and Systematic Sampling

If the population consists of linear trend and given by

$$X_i = i\,; \quad i = 1, 2, 3, \ldots, N$$

Then, $\sum_{i=1}^{N}X_i = \sum_{i=1}^{N}i = \frac{N(N+1)}{2}$

$$\sum_{i=1}^{N}X_i^2 = \sum_{i=1}^{N}i^2 = \frac{N(N+1)(2N+1)}{6}$$

Therefore,

$$\overline{X} = \frac{N+1}{2}$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}_{..}\right) = \frac{1}{(N-1)}\left[\sum_{i=1}^{N} X_i^2 - N\overline{X}^2\right]$$

$$S^2 = \frac{1}{N-1}\left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4}\right]$$

$$S^2 = \frac{N(N+1)}{2(N-1)}\left[\frac{(2N+1)}{3} - \frac{(N+1)}{2}\right]$$

$$= \frac{N(N+1)}{12}$$

$$\text{Var}\left(\overline{x}_{srswor}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2$$

$$= \frac{N-n}{N.n} \cdot \frac{N(N+1)}{12}$$

$$= \frac{nk-n}{n^2k} \cdot \frac{nk(nk+1)}{12}$$

$$\text{Var}\left(\overline{x}_{srswor}\right) = \frac{(k-1)(nk+1)}{12} \quad \ldots (1)$$

We have $S^2 = \frac{N(N+1)}{12}$ for population of N units.

In stratified random sampling we have

$$\text{Var}\left(\overline{x}_{st}\right) = \sum_{i=1}^{k} W_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2$$

In our case, there are n strata of size k and we draw one unit from each stratum so we put $N_i = k$, $n_i = 1$, $k = n$ and $W_i = N_i / N = 1/n$.

$$\text{Var}\left(\overline{x}_{st}\right) = \sum_{j=1}^{n} \frac{1}{n^2}\left(1 - \frac{1}{k}\right) S_i^2$$

Since i^{th} stratum consists of k units, we have

$$S_i^2 = \frac{k(k+1)}{12}$$

Therefore, $$\text{Var}\left(\overline{x}_{st}\right) = \frac{(k-1)}{n^2k} \cdot \frac{nk(k+1)}{12}$$

$$\text{Var}\left(\overline{x}_{st}\right) = \frac{\left(k^2-1\right)}{12n} \quad \ldots (2)$$

For finding out $\mathrm{Var}(\bar{x}_{sys})$, we have

$$\bar{x}_i = \text{mean of the values of } i^{th} \text{ sample} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$$

$$= \frac{1}{n}\left[i + (i+k) + (i+2k) + ... + (i+(n-1)k)\right]$$

$$= \frac{1}{n}\left[ni + \{1+2+3+...+(n-1)\}\, k\right]$$

$$= \frac{1}{n}\left[ni + \frac{(n-1)\,n}{2}.k\right] = i + \frac{(n-1)}{2}k$$

Also $\bar{x}_{..} = \bar{x} = \frac{N+1}{2} = \frac{nk+1}{2}$

$$\bar{x}_i - \bar{x}_{..} = i + \frac{(n-1)\,k}{2} - \frac{(nk+1)}{2}$$

$$= i - \frac{(k+1)}{2}$$

$$\mathrm{Var}(\bar{x}_{sys}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{x}_i - \bar{x}_{..})^2$$

$$= \frac{1}{k}\sum_{i=1}^{k}\left(i - \frac{(k+1)}{2}\right)^2$$

$$= \frac{1}{k}\sum_{i=1}^{k}\left(i^2 + \left(\frac{k+1}{2}\right)^2 - 2i\left(\frac{k+1}{2}\right)\right)$$

$$= \frac{1}{k}\sum_{i=1}^{k} i^2 + \left(\frac{k+1}{2}\right)^2 - \left(\frac{k+1}{k}\right)\sum_{i=1}^{k} i$$

$$= \frac{(k+1)(2k+1)}{6} + \frac{(k+1)^2}{4} - \frac{(k+1)^2}{2}$$

$$\mathrm{Var}(\bar{x})_{sys} = \frac{k^2-1}{12} \qquad \ldots (3)$$

From equations (1), (2) and (3), we get

$$\mathrm{Var}(\bar{x}_{st}) : \mathrm{Var}(\bar{x}_{sys}) : \mathrm{Var}(\bar{x}_{srswor})$$

$$\frac{(k^2-1)}{12n} : \frac{(k^2-1)}{12} : \frac{(k-1)(nk+1)}{12}$$

$$\frac{(k+1)}{n} : (k+1) : nk+1$$

$$\frac{1}{n} \;:\; 1 \;:\; n \text{ (approx)}$$

$$\mathrm{Var}(\bar{x}_{st}) \le \mathrm{Var}(\bar{x}_{sys}) \le \mathrm{Var}(\bar{x}_{srswor})$$

Example 1: In a class of Statistics, total number of students is 30. Select a systematic random sample of 10 students.The age of 30 students is given below:

Age:	22	25	22	21	22	25	24	23	22	21
	20	21	22	23	25	23	24	22	24	24
	21	20	23	21	22	20	20	21	22	25

Solution: We have given a population of size N = 30 values of the age of 30 students. Now, first of all we arrange all the values together with their serial numbers. Therefore,

Sr No:	1	2	3	4	5	6	7	8	9	10
Age:	22	25	22	21	22	25	24	23	22	21
Sr No:	11	12	13	14	15	16	17	18	19	20
Age:	20	21	22	23	25	23	24	22	24	24
Sr No.	21	22	23	24	25	26	27	28	29	30
Age:	21	20	23	21	22	20	20	21	22	25

After that we obtain value k as

$$k = \frac{N}{n} = \frac{30}{10} = 3$$

From first k values i.e. 1 to 3, we select a value randomly. Let we select the age 25 which is in 2^{nd} place in data. Therefore, our 1^{st} unit which selected in the sample is i = 25.

Now rest of the 9 units we will select systematically which are at the position (i+1k), (i+2k), (i+3k), …, (i+(n-1)k) in the given data. So according to the given data rest of the 9 units in our case would be the age given in 5^{th}, 8^{th}, 11^{th}, 14^{th} , 17^{th}, 20^{th}, 23^{rd}, 26^{th}, 29^{th} position.

Therefore, all the 10 units which has been selected in the sample are {25, 22, 23, 20, 23, 24, 24, 23, 20, 22}.

E1) The information regarding production of wheat (in Thousand kg) in 25 districts are collected, for a particular season. Select a systematic random sample of 7 units from the data given below:

23, 20 ,30, 37, 76, 36, 13, 36, 16, 58, 53, 83, 10, 15,13, 17, 12, 16, 17, 21, 20, 18, 61, 31, 71.

E2) A data of 50 values of heights (in cm) according to the roll no. of the students is given as follows:

146,	156,	152,	167,	178,	180,	172,	162,	148,
153,	161,	173,	163,	164,	175,	168,	161,	180,
173,	185,	169,	167,	168,	173,	145,	153,	154,
162,	164,	170,	172,	160,	161,	158,	152,	163,
165,	170,	168,	158,	149,	155,	160,	150,	149,
167,	176,	169,	159,	160.				

Select a systematic random sample of size 10.

4.5 INTRODUCTION TO CLUSTER SAMPLING

The population has been considered as a group of a finite number of distinct and identified units defined as sampling units. The smallest identity content in a population is known as element or elementary unit of the population. A group of such elementary units is known as cluster. Clusters are generally made up of which all the elements tend to have similar characteristics. When these clusters are treated as sampling units and few of them are selected either by equal or unequal probabilities then this procedure is known as cluster sampling. All the elements in selected clusters are to be observed, measured and interviewed. The number of elements in the cluster should be small and the number of clusters in the population should be large. For example, if we are interested in obtaining the information or data for monthly average income in a colony, then the whole colony may be divided into N numbers of block known as clusters and a simple random sample of n blocks is to be drawn. The individuals living in the selected clusters would be determined for interviewing to collect the information.

4.5.1 Significance of Cluster Sampling

Following are the various reasons, which cause problems in the selection of a sample of elementary units and cluster sampling enables us to overcome those problems:

1. When the sampling frame is unavailable, so the identifying and interviewing of sampling units is costly in terms of money, time consuming and need much labor. For example a list of households in metro city, list of farmhouse owners in a state, etc.;
2. The location of the identified sampling units may be situated far apart from one another and consume a lot of time and money to survey them;
3. It may not be possible to find well identifiable and easily locatable elementary units.

Thus, to overcome the above problems, cluster sampling yields satisfactory results in sampling of elementary units. The elementary units are formed in groups on the basis of location, class or area in cluster sampling.

4.5.2 Formation of Clusters

Certain precautions should be taken necessarily while dealing with the clusters sampling, which are given as follows:

1. The clusters should be made like that each elementary units should belong to one and only one cluster;
2. All units of similar characteristics should belong to the same cluster;
3. Each and every unit of the population should be included in any of the clusters constituting the population. In other words, there should neither be overlapping clusters nor omission of units;
4. All clusters should be heterogeneous themselves; and
5. Clusters should be as small as possible.

4.6 PROPERTIES OF CLUSTER SAMPLING

Notations

N = Number of clusters in the population

M = Number of elements in the clusters

n = Number of clusters in the sample

X_{ij} be the value of characteristic under study for the j^{th} element (j=1, 2,…, M) in the i^{th} cluster (i = 1, 2,…, N) of a population.

$\overline{X}_i = \sum_{j=1}^{M} X_{ij}/M$ is the mean of the i^{th} cluster in the population

$\overline{X} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} X_{ij}$ is the mean per element in the population

Similarly, x_{ij} be the value of characteristic under study for the j^{th} element (j =1, 2, …, M) in the i^{th} cluster (i = 1, 2,…, n) in the sample

Let $\overline{x}_i = \frac{1}{M}\sum_{j=1}^{M} x_{ij}$ is the mean of the i^{th} cluster in sample

and $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} \overline{x}_i$ is the mean of cluster means in the sample of size n

We have

$S_i^2 = \frac{1}{(M-1)}\sum_{j=1}^{M}(X_{ij} - \overline{X}_i)^2$ is the mean squar between elements within the i^{th} cluster

$S_w^2 = \frac{1}{N}\sum_{i=1}^{N} S_i^2$ the mean square within clusters

and $S_b^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(\overline{X}_i - \overline{X})^2$ is the mean squar between the cluster means in the population.

Therefore,

$S^2 = \frac{1}{(NM-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij} - \overline{X})^2$ denotes the mean square between elements in the population and

$$\rho = \frac{E(X_{ij} - \overline{X})(X_{ik} - \overline{X})}{E(X_{ij} - \overline{X})^2} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j=1}^{M}(X_{ij} - \overline{X})(X_{ik} - \overline{X})}{(M-1)(NM-1)S^2}$$

denote the intra-cluster correlation coefficient.

Theorem 2: In simple random sample without replacement of n clusters each containing M elements drawn form a population of N clusters, the sample mean $\overline{x}_n$ is an unbiased estimator $\overline{X}$ and its variance is given by

$$Var(\overline{x}_n) = \left(\frac{1-f}{n}\right)S_b^2 \cong \left(\frac{1-f}{n}\right)S_b\left[1+(M-1)\rho\right]$$

where, ρ is the intra-cluster correlation coefficient.

Proof: We have

$$E(\bar{x}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_i\right)$$

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

$$= \frac{1}{n}\sum_{i=1}^{N}\frac{n}{N}\bar{X}_i = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}X_{ij}$$

$$= \bar{X}$$

and now we have

$$Var(\bar{x}_n) = \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2$$

$$= \left(\frac{1-f}{n}\right)\sum_{i=1}^{N}\frac{(\bar{X}_i-\bar{X})^2}{(N-1)}$$

where f = n/N.

Using

$$\sum_{i=1}^{N}(\bar{X}_i-\bar{X})^2 = \sum_{i=1}^{N}\left(\frac{1}{M}\sum_{j=1}^{M}X_{ij}-\bar{X}\right)^2 = \frac{1}{M^2}\sum_{i=1}^{N}\left(\sum_{j=1}^{M}X_{ij}-M\bar{X}\right)^2$$

$$= \frac{1}{M^2}\sum_{i=1}^{N}\left(\sum_{j=1}^{M}(X_{ij}-\bar{X})\right)^2$$

$$\sum_{i=1}^{N}(\bar{X}_i-\bar{X})^2 = \frac{1}{M^2}\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij}-\bar{X})^2 + \frac{1}{M^2}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j=1}^{M}(X_{ij}-\bar{X})(X_{jk}-\bar{X})$$

$$= \frac{(NM-1)}{M^2}S^2[1+(M-1)\rho]$$

Therefore,

$$Var(\bar{x}_n) = \left(\frac{1-f}{n}\right)\left[\frac{(NM-1)}{M^2(N-1)}S^2\{1+(M-1)\rho\}\right]$$

$$= \left(\frac{1-f}{nM}\right)S^2[1+(M-1)\rho] \text{ for large } N$$

Variance in cluster sampling depends on the number of clusters in the sample, the size of the cluster, the intra-cluster correlation coefficient ρ and the mean square between the elements in the population. The variance of cluster sampling reduces to the variance of simple random sampling if M = 1.

Example 2: To determine the yield rate of wheat in a district of Punjab, 6 groups were constructed of 6 plots each. The data is given in the following table:

Plot No.	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
1	8	6	18	13	17	12
2	13	5	8	7	15	15
3	11	16	6	13	10	11
4	26	5	10	6	21	17
5	13	16	16	7	20	8
6	31	5	20	2	25	10

Select a cluster sample of 3 clusters from the given data and find sample mean.

Solution: In the given data, 6 groups have been formed and we have to draw a sample of 3 groups. Therefore, there will be 20 possible samples of size 3 which may be drawn from the population of size 6. Let we consider one sample of 3 groups drawn from population of 6 groups is {Group 1, Group 3 and Group 5}. Therefore, the sample mean will be

$$\bar{x} = \frac{n_1\bar{x}_1 + n_3\bar{x}_3 + n_5\bar{x}_5}{n_1 + n_3 + n_5}$$

where,

$$\text{Mean of Group 1} \qquad \bar{x}_1 = \frac{8+13+11+26+13+31}{6} = \frac{102}{6} = 17$$

$$\text{Mean of Group 3} \qquad \bar{x}_3 = \frac{18+8+6+10+16+20}{6} = \frac{78}{6} = 13$$

$$\text{Mean of Group 5} \qquad \bar{x}_5 = \frac{17+15+10+21+20+25}{6} = \frac{108}{6} = 18$$

Therefore,

$$\text{Grand Mean } \bar{x} = \frac{6\times17 + 6\times13 + 6\times18}{18}$$

$$= \frac{6\times(17+13+18)}{18} = \frac{6\times48}{18}$$

$$= 16$$

Hence, the sample mean is 16.

E3) A Housing board is constructing the 10 Duplex on each of 5 different locations. The plot areas (in square fit) as given in the following table:

S. No.	location 1	location 2	location 3	location 4	location 5
1	800	800	1300	2100	700
2	1300	700	1000	1600	2000
3	600	1500	1100	1300	1700
4	1800	1000	1600	600	1300
5	1300	1600	500	700	1600
6	1700	1300	1000	800	1300
7	1200	1800	1300	1200	1000
8	1300	600	1000	1500	500
9	500	700	1700	1600	1500
10	1000	1000	1000	1000	1000

Select a sample of 2 clusters with the help of cluster sampling scheme and calculate the sample mean.

4.7 INTRODUCTION TO TWO-STAGE SAMPLING

A sample survey as pointed out in Unit 1 of this block has certain limitations, mainly regarding the budget and time availability. Hence in survey too many elementary units is often not possible. In stratified random sampling, a sample is selected of optimum size from each stratum and then each and every unit selected from different stratum is to be observed. In Section 4.5 of this unit we have discussed cluster sampling in which the population is divided into some number of clusters and clusters were considered as sampling units. All the units in the selected clusters are enumerated completely. It has been pointed out there that cluster sampling is economically better than other samplings but the method restricts the spread of sample over the population which increased the variance of the estimator.

Instead of enumerating all the sampling units in the selected clusters, one can select a subsample of identified and specific units from the selected clusters by the same or different sampling methods.

The sampling which consists of selected clusters and then select the specified number of units from each selected cluster is known as two stage sampling. In this sampling technique, clusters being termed as primary stage units and units within clusters as secondary stage units. This method can be generalized upto three or more stages and is termed as multi stage sampling.

When large scale surveys on district, state or national level are to be conducted, it is one of the most suitable sampling schemes. For example, suppose we would like to draw the information about the monthly income of a household in a colony, it is better to select a sample of some blocks or wards and then households from the selected blocks or wards. In this procedure, the wards or blocks are the first stage units and the households are the second stage units.

4.7.1 Terminologies

N = Total number of first stage units.

n = Sample size of first stage units.

M = Number of second stage units in each first stage unit.

m = Number of second stage units in the sample from each first stage unit.

X_{ij} = Observation on the j^{th} second stage unit belonging to i^{th} first stage unit.

$\overline{X}_i$ = Mean of the i^{th} first stage unit $= \frac{1}{M}\sum_{j=1}^{M} X_{ij}$

$\bar{x}_i$ = Sample mean of the i^{th} first stage unit $= \frac{1}{m}\sum_{j=1}^{m} x_{ij}$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} \bar{x}_i$ = Overall sample mean on the basis of each second stage unit when sub – sampling has been done

$\bar{\bar{X}} = \frac{1}{N}\sum_{i=1}^{N} \bar{X}_i$ = Overall population Mean

4.8 PROPERTIES OF TWO-STAGE SAMPLING

Theorem 3: If the n first stage units and m second stage units from each selected first stage unit are selected by simple random sampling without replacement. There sample mean ($\bar{\bar{x}}$) is an unbiased estimator of population mean ($\bar{\bar{X}}$) and having the variance

$$Var(\bar{\bar{x}}) = \frac{(N-n)}{N}\frac{S_b^2}{n} + \frac{M-m}{M}\frac{S_w^2}{mn}$$

where, $S_b^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(\bar{X}_i - \bar{\bar{X}})^2$ and $S_w^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij} - \bar{X}_i)^2$

Proof: Since the units are selected in two stages by considering a probability sampling in each stage in two stage sampling. At both stages, selection procedures are to be considered in deriving expected value and variance of the sample statistic based on the number of units selected in second stage.

For getting the expected value and variance we have to follow:

$$E(\bar{\bar{x}}) = E_1[E_2[\bar{x}_i / i]] \quad \ldots (4)$$

$$Var(\bar{\bar{x}}) = Var_1[E_2(\bar{\bar{x}}/i)] + E_1[Var_2(\bar{\bar{x}}/i)] \quad \ldots (5)$$

Where E_1 and Var_1 are expectation and variance over the first stage units and E_2 and Var_2 are the conditional expectation and variance over the second stage unit for a given sample of first stage units.

Now since we draw second stage units from the first stage units by SRSWOR so

$$E_2[\bar{x}/i] = \bar{X}_i$$

Therefore, from equation (4), we have

$$E[\bar{\bar{x}}] = E_1[\bar{X}_i] = \bar{\bar{X}}$$

It shows that sample mean of all elements in the sample is an unbiased estimate of the population mean.

To obtain the variance

$$Var(\bar{\bar{x}}) = Var_1[E_2(\bar{\bar{x}}/i)] + E_1[Var_2(\bar{\bar{x}}/i)]$$

$$= \text{Var}_1\left(\overline{X}_i\right) + E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2\right]$$

$$= \frac{N-n}{Nn}S_b^2 + \frac{(M-m)}{mM}\frac{S_w^2}{n}$$

where, $S_w^2 = \frac{1}{N}\sum_{i=1}^{N}S_i^2$

Example 3: Select a sample of size 6 from given population of 36 units. The data given in 'Example 2' is divided in 6 clusters or groups each of them having 6 units.

Solution: Let us select 3 groups as first stage units from the given 6 groups. Let the selected units are Group-1, Group-3 and Group-6. Therefore the first stage sample is:

S. No.	Group-1	Group-3	Group-6
1	8	18	12
2	13	8	15
3	11	6	11
4	26	10	17
5	13	16	18
6	31	20	10

Now, from the selected first stage units, we shall select the 6 second stage units. These units are selected on the basis of their importance. Let we select 2 units each from these three selected first stage units of size 6.

Let us select 2nd & 4th second stage units from Group-1, 1st & 4th unit from Group-3 and 3rd & 6th unit from Group-6.

Therefore, the second stage sample units selected from 3 first stage units of size 6 are {13, 26, 18, 10, 11, 10 }.

Let us answer the given exercise.

E4) Select a first stage sample of size 2 then the second stage sample of size 10, from the data given in E3) by two-stage sampling method.

4.9 SUMMARY

In this unit, we have discussed:

1. Systematic random sampling;
2. How to draw a systematic random sample and estimate the population mean;
3. The variance of the estimate of the population mean;
4. The cluster sampling and the estimate of the variance of the sample mean; and
5. Method of drawing a two-stage sample; and
6. Method of solving the numerical examples

4.10 SOLUTIONS / ANSWERS

E1) First of all we arrange the data of production of wheat (in Thousand kg) with their serial order:

Sr.No. :	1	2	3	4	5	6	7	8	9	10	11	12	13
Production:	23	20	30	37	76	36	13	36	16	58	53	83	10
Sr. No.:	14	15	16	17	18	19	20	21	22	23	24	25	
Production:	15	13	17	12	16	17	21	20	18	61	31	71	

Now we obtain a number $k = \frac{N}{n} = \frac{25}{7} = 3.5$

We have N = 25 and n = 7

So we have to take k = 4

Now, from first 4 values in serially arranged data let we select 3rd value (i =30), so this will be our 1st sample unit selected in the sample by random method. Now remaining 6 values will be selected in systematic way i.e. $(i+1k)^{th}$, $(i+2k)^{th}$, …, $(i+6k)^{th}$ order value in the data. So in this way, we have to select the values which are at 7^{th}, 11^{th}, 15^{th}, 19^{th}, 23^{rd} and 27^{th}position. But in the data, only 25 values are available. So, we have to adopt the circular systematic sampling method for the selection of all 7 units. By following the circular systematic sample after selecting the first unit ($i = 3^{rd}$) in between 1 to 25 the remaining units would be 7^{th}, 11^{th}, 15^{th}, 19^{th}, 23^{rd} and 2^{nd} positioned units.

Therefore, from the population of 25 units the systematic random sample of size 7 would be {30, 13, 53, 13, 17, 61, 20}.

E2) We have N = 50 and we have to draw a systematic sample of size 10. We arrange the values of the heights of 50 students corresponding to their roll numbers from 1 to 50.

Roll No:	1	2	3	4	5	6	7	8	9
Height:	146	156	152	167	178	180	172	162	148
Roll No:	10	11	12	13	14	15	16	17	18
Height:	153	161	173	163	164	175	168	161	180
Roll No:	19	20	21	22	23	24	25	26	27
Height:	173	185	169	167	168	173	145	153	154
Roll No:	28	29	30	31	32	33	34	35	36
Height:	162	164	170	172	160	161	158	152	163
Roll No:	37	38	39	40	41	42	43	44	45
Height:	165	170	168	158	149	155	160	150	149
Roll No:	46	47	48	49	50				
Height:	167	176	169	159	160				

Now we shall obtain a number $k = \frac{N}{n} = \frac{50}{10} = 5$ that is k = 5. So we have to select the first unit of the sample of size 10 by random selection method from serially 1 to 5. Let we have selected the 4^{th} unit in order. So the 1^{st}unit selected in the sample is 167.

Now we shall select the remaining 9 unit in a systematic way. We have
i = 4 and k = 5 so the remaining 9 sample unit to be selected in the sample would be 9^{th},14^{th}, 19^{th}, 24^{th} , 29^{th} , 34^{th}, 39^{th}, 44^{th} and 49^{th} order in the data.

Therefore, the values of systematic random sample of size 10 selected from the population of size 50 would be 167, 148, 164,173, 173, 164, 158, 168, 150, 159.

E3) In cluster sampling, we know that if population is divided in several N homogeneous groups (known as clusters) then we have to select the clusters as sample units for selection of sample of size n.

Here, in the given data N = 5 and n =2. Therefore, there will be ${}^{N}C_{n} = {}^{5}C_{2} = 10$ number of possible sample of size 2 clusters.

Let us select the Location 2 and Location 4 as the sample units. Therefore, the sample of 2 locations selected from population of 5 locations is

Location	Size of Plots										Total
Location 2	800	700	1500	1000	1600	1300	1800	600	700	1000	11000
Location 4	2100	1600	1300	800	1000	900	1200	1500	1600	1000	13000

Therefore, sample mean of first sample unit

$$\bar{x}_1 = \frac{11000}{10} = 1100$$

and sample mean of second sample unit

$$\bar{x}_2 = \frac{13000}{10} = 1300$$

Therefore, the sample mean

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 +}{n_1 + n_2} = \frac{10\times1100 + 10\times1300}{10+10}$$

$$= \frac{10(1100+1300)}{20} = 1200 \quad \text{square fit}$$

E4) The data given in E3) are having 50 number of sizes of plots, which are divided into 5 groups on the basis of the different locations, i.e. Location 1, Location 2, …, Location 5. Now, we select a first stage sample of two locations. Let it be Location 2 and Location 5 as first stage sample units.

Therefore, we have

Plot No.	Location 2	Location 5
1	800	700
2	700	2000
3	1500	1700
4	1000	1300
5	1600	1600
6	1300	1300
7	1800	1000
8	600	500
9	700	1500
10	1000	1400
Total	11,000	13,000

Now, from the selected first stage units we shall select a second stage sample of 10 units by selecting 5 units each from both locations. Let the 2^{nd}, 3^{rd}, 6^{th}, 7^{th} and 9^{th} are selected from the Location 2 and 3^{rd}, 5^{th}, 7^{th}, 8^{th} and 10^{th} unit are selected from the Location 5.

Therefore, the sample selected by the two-stage sampling method is {700, 1500, 1300, 1800, 700, 1700, 1600, 1000, 500, 1400}.

Indira Gandhi National Open University
School of Sciences

MST - 005
Statistical Techniques

Block

2

ANALYSIS OF VARIANCE

UNIT 5
Introduction to Analysis of Variance **5**

UNIT 6
One-way Analysis of Variance **21**

UNIT 7
Two-way Analysis of Variance **41**

UNIT 8
Two-way Analysis of Variance with m Observations per Cell **61**

Curriculum and Course Design Committee

Prof. K. R. Srivathsan
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Parvin Sinclair
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Geeta Kaicker
Director, School of Sciences
IGNOU, New Delhi

Prof. R. M. Pandey
Department of Bio-Statistics
All India Institute of Medical Sciences
New Delhi

Prof. Jagdish Prasad
Department of Statistics
University of Rajasthan, Jaipur

Prof. Rahul Roy
Maths and Stat. Unit
Indian Statistical Institute, New Delhi

Dr. Diwakar Shukla
Department of Mathematics and Statistics
Dr. Hari Singh Gaur University, Sagar

Prof. G. N. Singh
Department of Applied Mathematics
I S M Dhanbad

Prof. Rakesh Srivastava
Department of Statistics
M. S. University of Baroda, Vadodara

Dr. Gulshan Lal Taneja
Department of Mathematics
M. D. University, Rohtak

Faculty Members, School of Sciences, IGNOU

Statistics
Dr. Neha Garg
Dr. Nitin Gupta
Mr. Rajesh Kaliraman
Dr. Manish Trivedi

Mathematics
Dr. Deepika Garg
Prof. Poornima Mital
Prof. Sujatha Varma
Dr. S. Venkataraman

Block Preparation Team

Content Editor
Prof. G. K. Shukla
Decision Science Group
Indian Institute of Management, Lucknow

Language Editor
Dr. Parmod Kumar
School of Humanities, IGNOU

Secretarial Support
Mr. Deepak Singh

Course Writer
Prof. Jagdish Prasad
Department of Statistics
University of Rajasthan, Jaipur

Formated By
Dr. Manish Trivedi
Mr. Prabhat Kumar Sangal
School of Sciences, IGNOU

Programme and Course Coordinator: Dr. Manish Trivedi

Block Production

Mr. Y. N. Sharma, SO (P.)
School of Sciences, IGNOU

Acknowledgement: We gratefully acknowledge Prof. Geeta Kaicker, Director, School of Sciences for her great support and guidance.

December, 2011

ISBN – 978-81-266-5785-8

Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by Director, School of Sciences.

Laser Type set by: Rajshree Computers, V-166A, Bhagwati Vihar, (Near Sector-2, Dwarka), Uttam Nagar, New Delhi-110059

Printed at: Gita Offset Printers Pvt. Ltd., C-90, Okhla Industrial Area, Phase-I, New Delhi-110020.

ANALYSIS OF VARIANCE

The impact of the ever increasing number of applications of analysis of variance in the field of Physical Sciences, Life Sciences, Horticulture, Agriculture, Engineering, Management, Medical Sciences, Pharmaceutical and Social Science makes it indispensable that uses of these disciplines are offered a basic training and knowledge of Statistics specially analysis of variance. This block on analysis of variance is designed in such a way to serve as a basic course on text, meeting the sufficient requirements of a learner or researcher of the above discipline.

It is a fact that before understanding the concept of analysis of variance, one has to have some idea or basic knowledge of Statistics and procedure of estimation and testing of hypothesis. This course is basically designed for applied purpose. So, each concept is introduced in an intuitive way. The relevance of all the procedures and techniques is derived by considering different types of examples. The aim of this course is to enable you to successfully handle and solve any statistical problem to analysis of variance in your field or discipline. We have included many practical exercises and examples, solved and unsolved, on various places in the course.

This block is divided into four units. In Unit 5, we introduced the concept of analysis of variances, its definition, terminology used, assumptions of analysis of variance and its uses. The concept of linear models in analysis of variance technique is also explained in this unit.

In Unit 6, one-way classified data with assumptions and its analysis is explained. The basic assumptions are explored and the expectation of various sums of squares is also derived in this unit.

The procedure for analysis of two-way classified data is described in Unit 7. We shall be dealing with analysis of variance technique in two-way classified data with m-observations per cell in Unit 8.

It is advised that the learners should do a lot of practice for commanding on the techniques and the exercises, given in the block, by using calculator or computer. If you are interested to learn more, you may look up or consider or consult some more books on the subject analysis of variance or you can see on the internet.

Suggested Readings:

- Goon, A. M., Gupta, M. K. and Das Gupta, B.; Fundamentals of Statistics, Vol II, World Press, Calcutta.
- Gupta, S. C. and Kapoor, V. K.; Fundamentals of Applied Statistics, Sultan Chand & Sons.
- Goulden, C. H.; Methods of Statistical Analysis (Ch. 5), Asia Publishing House, 1959.
- Guenther, W. C.; The Analysis of Variance, Prentice-Hall, 1964.
- Scheffe, H.; The Analysis of Variance (Chs. 3, 4, 7, 8), John Wiley, 1961.

Notations and Symbols

Y	: Response variable/ Dependent variable
X	: Explanatory variable/Independent variable/Predictor variable/Treatment/Factor/Effect
$\bar{y}$	: Mean of response variable
$\bar{x}$	: Mean of explanatory variable
y_{ij}	: j^{th} observation in the i^{th} level of a factor A
y_{ijk}	: k^{th} observation under i^{th} level of factor A and j^{th} level of factor B
μ	: An over all mean or Grand mean
μ_i	: Mean of i^{th} level of a factor A
α_i	: Effect of i^{th} level of a factor A
e_{ij}	: Error term
n_i	: Number of observations in i^{th} level of a factor
E	: Residual sum of squares
H_0	: Null hypothesis
H_1	: Alternative hypothesis
df/DF	: Degrees of freedom
$V(e_{ij})$	: Variance of error e_{ij}
F	: F- statistic / Variation ratio
α	: Level of significance
CD	: Critical difference
β_j	: Effect due to j^{th} level of factor B
$(\alpha\beta)_{ij}$	: Interaction effect between i^{th} level of factor A and j^{th} level of factor B
p	: Number of levels of factor A
q	: Number of levels of factor B
SSA	: Sum of Squares due to factor A
SSB	: Sum of Squares due to factor B
TSS	: Total Sum of Squares
SST	: Sum of Squares due to Treatments or different levels of a factor
SSE	: Sum of Squares due to Error or Residual
$N(\mu,\sigma^2)$	: Normally distributed with mean (μ) and variance (σ^2)
i.i.d.	: Independently and identically distributed.
RSS	: Raw Sum of Squares
CF	: Correction Factor
SS	: Sum of Squares
MSS	: Mean Sum of Squares
SV	: Source of Variation

UNIT 5 INTRODUCTION TO ANALYSIS OF VARIANCE

Structure

5.1 Introduction
Objectives
5.2 Analysis of Variance
5.3 Basic Definitions of the Terms used in Analysis of Variance
5.4 Basic Assumptions in Analysis of Variance
Assumptions of Randomness
Assumptions of Additivity
Equality of Variances or Homoscedasticity and Zero Correlation
Assumptions of Normality
5.5 Linear Models Used in Analysis of Variance
Fixed Effect Model
Random Effect Model
Mixed Effect Model
5.6 Uses of ANOVA
5.7 Summary
5.8 Solutions /Answers

5.1 INTRODUCTION

As its name suggests, the analysis of variance focuses on variability. It involves the calculation of several measures of variability, all of which comes down to one or another version of the measure of variability such as the sum of squared deviations or mean sum of squared deviations. The statistical technique known as "Analysis of Variance", commonly referred to by the acronym ANOVA was developed by Professor R. A. Fisher in 1920's. Variation is inherent in nature, so analysis of variance means examining the variation present in data or parts of data. In other words, analysis of variance means to find out the cause of variation in the data. The total variation in any set of numerical data of an experiment is due to number of causes which may be (i) Assignable causes; and (ii) Unassignable / Chance causes.

The variation in the data due to assignable causes can be detected, measured and controlled whereas the variation due to chance causes is not in the control of human being and cannot be traced or find out separately.

The reason, this analysis is called analysis of variance rather than multi-group mean analysis (or something like that), is because it compares group means by analysing comparisons of variance estimates. Analysis of variance facilitates the analysis and interpretation of data from field trials and laboratory experiments in agriculture and biological research. Today, it constitutes one of the principal research tools of the biological scientists, and its use is spreading rapidly in the social sciences, the physical sciences, in the engineering, in management, etc.

In Unit 11 of MST-004, we compared means from two independent groups by using t-test. But if we are interested to test more than two independent groups then t-test cannot be applied and firstly we have to apply analysis of variance technique. An F-test is used to test the means of several groups. This F-test was named 'F' in honor of Professor R. A. Fisher by G. W. Snedecor. ANOVA is helpful because it possesses an advantage over a two sample t-test. The multiple two sample t-test would result in an increase of chance of committing a type I error.

The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between two population means. In an situation when we have more than two population to consider at a time and want to test the means of these population are same. For example, five doses of a drug are applied to four patients each and responses / values of dependent variable (observations) of these twenty patients are obtained. Now, we may be interested in finding out whether the effect of these five doses of drug on the patients is significantly differs. The answer to this problem is provided by the technique of analysis of variance. Thus, the analysis of variance technique is used to test the homogeneity of several population means.

Thus, the analysis of variance technique is a powerful statistical tool for tests of significance in comparing more than two means.

The notations and basic definitions of the important terms are provided in Section 5.2, so that you would become familiar with the basic terminologies, notations and the various types of analysis of variance technique. Section 5.3 describes the various assumptions involved in analysis of variance whereas Section 5.4 explains the various types of linear models used in analysis of variance. Applications of analysis of variance are explored in Section 5.5.

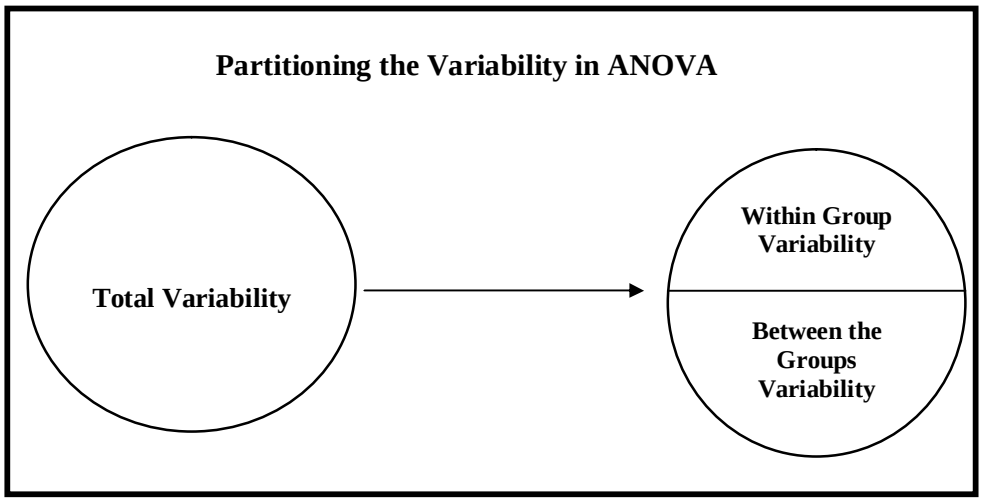
Objectives

After reading this unit, you would be able to

- become familiar with the analysis of variance technique;
- describe the various types of analysis of variance;
- describe the various assumptions involved in analysis of variance;
- define the various types of linear models used in analysis of variance; and
- describe the applications of analysis of variance.

5.2 ANALYSIS OF VARIANCE

According to Professor R. A. Fisher, Analysis of Variance (ANOVA) is "Separation of variance ascribable to one group of causes from the variance ascribable to other group". So, by this technique, the total variation present in the data are divided into two components of variation one is due to assignable causes (between the groups variability) or other is variation due to chance causes (within group variability).



The analysis of variance technique solves the problems of estimating and testing to determine, whether to infer the existence of true difference among "treatment" means, among variety means and under certain conditions among other means with respect to the problem of estimation. The analysis of variance is simply the form of the method of least squares discussed in Unit 5 of MST-002. Analysis of variance technique can be classified as follows:

1. Parametric ANOVA.
2. Non Parametric ANOVA.

Parametric ANOVA can be classified as simply ANOVA if only one response variable is considered. If more than one response variables are under consideration than it is called multivariate analysis of variance (MANOVA).

If we consider, only one independent variable which affects the response / dependent variable then it is called One-way ANOVA. If the independent variables / explanatory variables are more than one i.e. n (say) then it is called n-way ANOVA. If n is equal to two than the ANOVA is called Two-way classified ANOVA.

Factorial ANOVA is used when the experimenter wants to study the interaction effects among the explanatory variables. Repeated measure ANOVA is used when the same subjects (experimental units) are used for each treatment (levels of explanatory variable). Multivariate analysis of variance (MANOVA) is used when there is more than one response variable.

The F test in ANOVA has been used with normality assumption. But there are some cases where the distribution of response variable and its transformation is not normal. In such cases, where the variable Y_{ij} is non-normal and the appropriate information which will make normal is unknown, we need non-parametric ANOVA methods. You can set Kruskal-Wallis One-way ANOVA by ranks. One-way repeated measures ANOVA for non-parametric is Friedman test among the explanatory variables. This technique can be used in different statistical linear models, like in Fixed effect model, Random effect model and Mixed effect model.

5.3 Basic Definitions of the Terms Used in Analysis of Variance

Once you have familiarized yourself with the terminology of analysis of variance, you will find it easier to grasp many of the parametric techniques that you read about in this unit. Some of the terms described below may be referred to by one of many names, as indicated below:

1. **Variable**
   A characteristic or attribute varies in a measurable way between subjects in a sample.

2. **Response/Dependent Variable Y**
   It describes the measurements, usually on a continuous scale, of the variable of interest (e.g. weight: what causes variation in weight?).

3. **Explanatory /Independent /Predictor Variable/Treatment/ Factor Effect X**
   Explanatory variable is controlled by experimenter, which may affect response. Response does not affect explanatory variable. The non-random measurements or observations (e.g. treatments fixed by experimental design), which are hypothesized in a statistical model to have predictive power over the response variable. This hypothesis is tested by calculating sums of squares and looking for a variation in Y between levels of X that exceeds the variation within levels. An explanatory variable can be categorical (e.g. sex, with 2 levels of male and female), or continuous (e.g. height with a continuum of possibilities). The explanatory variable is assumed to be 'independent' in the sense of being independent of the response variable: i.e. weight can vary with height, but height is independent of weight. The values of X are assumed to be measured precisely, without error, permitting an accurate estimate of their influence on Y. Explanatory variable may be fixed, random or mixed as per statistical model used.

4. **Variates/ Replicates/Observations/Scores/Data Points**
   The replicate observations of the response variable (Y_1, Y_2, …, Y_i, …, Y_N) are measured at each level of the explanatory variable. These are the data points, each usually obtained from a different subject to ensure that the sample size reflects N independent replicates (i.e. it is not inflated by non-independent data 'pseudo replication').

5. **Sample**
   The collection of observations measured at a level of X (e.g. body weights were measured from one sample of males and another of females to test the effect of Sex on Weight). If X is continuous, the sample comprises all the measures of Y on X (e.g. Weight on Height).

6. **Sum of Squares**
   The squared distance between each data point (Y_i) and the sample mean ($\overline{y}$), summed for all N data points. The squared deviations measure variation in a form which can be partitioned into different components that sum to give the total variation (e.g. the component of variation between samples and the component of variation within samples).

7. **Variance and Standard Deviation**
   The variance in a normally distributed population is described by the average of N squared deviations from the mean. Variance usually refers to a sample, however, in which case it is calculated as the sum of squares divided by (N−1) rather than N. Its positive root is then the standard deviation (SD) which describes the dispersion of normally distributed variables (e.g. 95% lying within 1.96 standard deviations of the mean when N is large).

8. **Statistical Model Y = f (X) + e, where f (X) is Some Function of X**
   It is a statement of the hypothesized relationship between the response variable and the predictor variable. A simple model would be: Weight = Sex + e. The '=' does not signify a literal equality, but a statistical dependency. So the statistical analysis is going to test the hypothesis that variation in the response variable on the left of the equals sign (Weight) is explained or predicted by the factor on the right (Sex), in addition to a component of random variation (the error term e). An analysis of variance will test whether significantly more of the variation in 'Weight' falls between the categories of 'male' and 'female', and so is explained by the independent variable 'Sex' that lies within each category (the random variation 'e'). The error term is often dropped from the model description though it is always present in the model structure, as the random variation against which to calibrate the variation between levels of X in testing for a significant explanation (the F-ratio).

9. **Null Hypothesis H_0**
   While a statistical model can propose a hypothesis, that response variable Y depends on independent variable X, the statistical analysis can only seek to reject a null hypothesis that Y does not vary with X. This is because it is always easier to find out how different things are than to know how much they are the same, so the statistician's easiest objective is to establish the probability of a deviation away from random expectation rather than towards any particular alternative. If the analysis reveals a sufficiently small probability that the null hypothesis is true, then we can reject it and state that Y evidently depends on X in some way or hypothesis of null hypothesis.

10. **One-Way ANOVA Y = f (X) + e**
    An analysis of variance (ANOVA) is used to test the model hypothesis that variation in the response variable Y can be partitioned into the different levels of a single explanatory variable X (e.g. Weight = Sex). If X is a continuous variable, then the analysis is equivalent to a linear regression, which tests for a significant slope in the best fit line describing change of Y with X (e.g. Weight with Height).

11. **Two-Way ANOVA $Y = X_1 + X_2 + X_1 X_2 + e$**
    Test of the hypothesis that variation in Y can be explained by one or both variables X_1 and X_2. If X_1 and X_2 are categorical and Y has been measured only once in each combination of levels of X_1 and X_2, then the interaction effect $X_1.X_2$ cannot be estimated. Otherwise a significant interaction term means that the effect of X_1 is modulated by X_2 (e.g. the effect of Sex (X_1), on Weight (Y) depends on Nationality (X_2). If one of the explanatory variables is continuous, then the analysis is equivalent to a linear regression with one line for each level of the categorical

variable (e.g. graph of -Weight by Height, with one line for males and one for females): different intercepts signify a significant effect of the categorical variable, different slopes signify a significant interaction effect with the continuous variable.

12. **Error/Residual**
    The amount by which an observed value of variable differs from the value predicted by the model is known as error. Errors or residuals are the segments of scores not accounted for by the analysis. In analysis of variance, the errors are assumed to be independent of each other, and normally distributed about the sample means. They are also assumed to be identically distributed for each sample (since the analysis is seeking only a significant difference between sample means), which is known as the assumption of homogeneity of variances.

13. **Normal Distribution**
    It is a bell-shaped frequency distribution of a continuous variable. The formula for the normal distribution contains two parameters: the mean, giving its location, and the standard deviation, giving the shape of the symmetrical 'bell'. This distribution arises commonly in nature when myriad independent forces, themselves subject to variation, combine additively to produce a central tendency. The technique of analysis of variance is constructed on the assumption that the component of random variation takes a normal distribution. This is because the various sum of squares that are used to describe variance in an ANOVA accurately reflect the true variation between and within samples only if the residuals are normally distributed about sample.

14. **Degrees of Freedom (df)**
    The F-ratio in an analysis of variance is always presented with two sets of degrees of freedom, the first corresponding to one less than a level (p) of the explanatory variable i.e. (p−1) and the second to the remaining error degrees of freedom i.e. (n−p) in one-way classified data.

15. **Variance Ratio / F-statistic**
    The statistic calculated by analysis of variance, which reveals the significance of the hypothesis that response variable depend on explanatory variable. It comprises the ratio of two mean squares i.e. Mean sum of squares due to explanatory variable divided by Mean sum of squares due to error. A large proportion indicates a signified effect of independent variable.

16. **Significance**
    This is the probability of mistakenly rejecting a null hypothesis that is actually true.

17. **P -Value**
    In biological sciences a critical value p = 0.05 is generally taken as marking an acceptable boundary of significance. A large F-Ratio signifies a small probability that the null hypothesis is true.

18. **Mean Sum of Squares**
    Mean sum of squares is the average sum of squares. In other words, “the sum of squares of deviations from the ‘mean X’ or ‘error e’ divided by its appropriate degrees of freedom”.

19. **Population**
    All subjects / experimental units possess a common characteristic that is being studied whereas a group of subjects selected from the target population is known as sample.

20. **Parameter and Statistic**
    Characteristics /measures obtained from a population are known as parameters whether the characteristics/measures obtained a sample are known as statistic.

E1) Describe the analysis of variance and differentiate between One-way and Two-way ANOVA.

5.4 BASIC ASSUMPTIONS IN ANALYSIS OF VARIANCE

The analysis of variance has been studied from several approaches, the most common of which is to use a linear model that relates the response variable to the treatments (explanatory variables). Even when the statistical model is non linear, it can be approximated by a linear model for which an analysis of variance may be appropriate.

When analysis of variance is used as a method of statistical inference for inferring properties of the "Population" from which the data are drawn (taken) then certain assumptions about the "Population" and the sampling procedure by means of which the data are obtained, must be fulfilled, if the inference are to be valid. So, ANOVA makes certain assumption about the nature of the experimental data that have to be at least approximately true before the method can be validly applied.

Before explaining the assumptions of ANOVA the response variable Y demonstrated in the following two-way table in which there are r-rows (varieties) and c-columns (treatments).

	1	2	...	j	...	c	Total	Row Means
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1c}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2c}	$y_{2.}$	$\bar{y}_{2.}$
.	.	.			.	.	.	.
.	.	.			.	.	.	.
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ic}	$y_{i.}$	$\bar{y}_{i.}$
.	.	.			.	.	.	.
.	.	.			.	.	.	.
r	y_{r1}	y_{r2}	...	y_{rj}	...	y_{rc}	$y_{r.}$	$\bar{y}_{r.}$
Total	$y_{.1}$	$y_{.2}$	...	$y_{.j}$	...	$y_{.c}$	$y_{..}$	
Column Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	...	$\bar{y}_{.j}$	...	$\bar{y}_{.c}$		$\bar{y}_{..}$

y_{ij} = is the value of the response variable Y occurring in the i^{th} row and j^{th} column.

$y_{i.} = \sum_{j=1}^{c} y_{ij}$ for all j =1, 2, …, c; = Total corresponding to the i^{th} row

$\bar{y}_{i.} = \frac{1}{c}\sum_{j=1}^{c} y_{ij}$ for all j = 1, 2,…, c; = Mean corresponding to the i^{th} row

$y_{.j} = \sum_{i=1}^{r} y_{ij}$ for all i =1, 2, …, r; = Total corresponding to the j^{th} column

$\bar{y}_{.j} = \frac{1}{r}\sum_{i=1}^{r} y_{ij}$ for all i = 1, 2,…, r; = Mean corresponding to the j^{th} column

$y.. = \sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}$ for all i=1, 2, …, r; j=1, 2, …, c; = Total of all the observations

$\bar{y}.. = \frac{1}{rc}\sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}$ for all i=1, 2, …, r; j=1,2, …, c; =Mean of all the observations

Following are the assumptions satisfied by the ANOVA technique:

5.4.1 Assumption of Randomness

The values y_{ij} are (observed values/response variable/dependent variable) random variables that distributed about so called true mean values (expected value) μ_{ij} (i=1, 2, ..., r; j=1,2, ..., c) are fixed constants.

In statistical language this assumption states that, of some particular type of experiment leading to value y_{ij} were repeated indefinitely, then the value y_{ij} would vary at random about an average value equal to μ_{ij}, which is therefore a parameter that characterizes the expected value of the y_{ij}.

From this assumption an unbiased estimator of any linear function of the μ_{ij} with known coefficient is obtained by the same linear function of the y_{ij}.

Further, if the variances of the y_{ij} about their respective means and their inter correlations are known, then the variances of any linear function of the y_{ij} can be evaluated and provides a measure of the precision of this linear function of the y_{ij} as an unbiased estimator of corresponding linear function of the μ_{ij}.

5.4.2 Assumption of Additivity

True means (μ_{ij}) are simple additive functions of the corresponding marginal means and the general mean, that is,

$$y_{ij} = \mu_{ij} + e_{ij}$$

$$\mu_{ij} = \mu_{..} + (\mu_{i.} - \mu_{..}) + (\mu_{.j} - \mu_{..})$$

or $y_{ij} = \mu_{..} + (\mu_{i.} - \mu_{..}) + (\mu_{.j} - \mu_{..}) + e_{ij}$ for (i=1, 2, ..., r; j=1, 2, ..., c)

then the statistical inferences that may be based upon the y_{ij} are of a much more satisfactory. Or we can say that the observed value/response variable y_{ij} can be the sum of three parts:

1. The overall mean of the observations.

2. Treatments/factors/classification effects.
3. A random element or error effect drawn from normally distributed populations. The random element reflects the combined effect of natural variation between observations and errors of measurement.

When the above additivity assumption is satisfied, the difference between arbitrary pair of row-wise marginal means, e.g. $\mu_{1.}$ and $\mu_{2.}$ (say) is a comprehensive measure of the average difference in effectiveness of the factors identified with these rows. When the above assumption does not satisfied then $\mu_{1.} - \mu_{2.}$ is merely a measure of the average difference between the effects of the corresponding row factors when the column factors are as in the experiment concerned. Similarly, the actual mean difference in effectiveness of a pair of column factors will depend upon row factors concerned. Hence, when additivity does not prevail, we say that there are interactions between row factor and column factor. Thus, additivity implies that the true mean yield of level of one factor is greater (or less) than the true mean of another level of a factor by an amount additive constant not a multiplier that is the same for each of the another factor concerned, and conversely the true mean yield with levels of another factor is greater (or less) than the true mean yield with levels of another factor by an amount that does not depend upon the previous factor concerned, which is exactly that there are no "Interaction" between first factor & second factor.

When assumptions 1 and 2 are satisfied that the difference between any pair of row wise means of the observations y_{ij} e.g., $\bar{y}_{1.} - \bar{y}_{3.}$ is an unbiased estimator of the general average difference in effectiveness of the row factors concerned i.e. of $\mu_{1.} - \mu_{3.}$. Similarly, for column wise means of the observations.

5.4.3 Equality of Variances or Homoscedasticity and Zero Correlation

The random variables y_{ij} are homoscedastic and mutually uncorrelated, that is, they have a common variance σ^2 although means vary from group to group. The variances should be constant in all groups and all covariance among them are zero so, it is desirable before using ANOVA technique the assumption of homogeneity of variance and normality should jointly be tested.

In general, it is not possible to derive from the observations y_{ij}, the unbiased estimates of variances of the y_{ij} and any particular linear functions of them, unless the assumptions 1, 2 and 3 are not satisfied.

If assumptions 1, 2 and 3 are all satisfied, an unbiased estimate of the variances of the differences of two observed row means and an unbiased estimate of the variances of the differences of two observed column means can be evaluated.

Suppose, assumption 1 and assumption 2 are satisfied and assumption 3 is not. In this case, four values of σ^2 would be the expected value of sum of square due to rows, due to column, due to residual and due to total, which may be denoted by σ_r^2, σ_c^2, σ_e^2 and σ_t^2. These will be the complex weighted means of the variances and covariances of the $y_{ij.}$

So, it is desirable before using ANOVA technique the assumption of homogeneity of variance should be tested. Although the means may vary

from group to group the variances should be constant in all the groups. The following procedures are widely used for this purpose:

Bartlett's Test for Homogeneity of Variances

The null hypothesis for equal variances is

$$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

against $H_1: \sigma_1^2 \neq \sigma_2^2 \neq \ldots \neq \sigma_k^2$

From each group of size n_i taken from i^{th} population, i = 1, 2, ... , k we can calculate S_1^2, S_2^2, ... , S_k^2 where,

$$S_i^2 = \frac{1}{n_i - 1}\sum_{i=1}^{k}\left(y_{ij} - \bar{y}_{i.}\right)^2 \qquad \text{for all } i = 1, 2, \ldots, k$$

Now, let $\nu_i = n_i - 1$; where, ν_i are the degrees of freedom of S_i^2

and $$\nu = \sum_{i=1}^{k}\nu_i \qquad \text{for all } i = 1, 2, \ldots, k$$

and $$S^2 = \sum_{i=1}^{k}\nu_i S_i^2$$

The Bartlett's test statistic M is defined by

$$M = \nu \log S^2 - \sum_{i=1}^{k}\nu_i \log S_i^2$$

When none of the ν_i's degrees of freedom is small, M is distributed as χ^2_{k-1}. The χ^2 approximation is generally acceptable if all n_i are at least 5. This is slightly biased test. So it can be improved by dividing M by the factor

$$C = 1 + \frac{1}{3(k-1)}\sum_{i=1}^{k}\left(\frac{1}{\nu_i} - \frac{1}{\nu}\right)$$

Instead of M, it is suggested to use M/C for the test statistic.

Levene's Test for Homogeneity of Variances

Levene's test offers a more robust alternative to Bartlett's procedure. That means it will be less likely to reject a true hypothesis of equality of variances just because the distribution of the sampled populations are not normal. When non normality is suspected, Levene's procedure is a better choice then Bartlett's goodness of fit test because the value of the response variance occurring in the row and column.

Levene's test is used to test if k samples have equal variances. Equal variances across samples are called homogeneity of variance. Some statistical tests, for example the analysis of variance, assume that variances are equal across groups or samples. The Levene's test can be used to verify that assumption.

Levene's test is an alternative to the Bartlett test. The Levene's test is less sensitive than the Bartlett test to departures from normality. If you have strong evidence that your data do in fact come from a normal, or nearly normal distribution, then Bartlett's test has better performance.

The Levene's test is defined as:

$$H_0: \sigma_1 = \sigma_2 = \ldots = \sigma_k$$

$H_1: \sigma_i \neq \sigma_j$ for at least one pair (i, j).

Given a variable Y (response variable) with sample of size N divided into k subgroups, where n_i is the sample size of the i^{th} subgroup, the Levene's test statistic is defined as:

$$W = \frac{(N-k)\sum_{i=1}^{k} n_i \left(\overline{Z}_{i.} - \overline{Z}_{..}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Z_{ij} - \overline{Z}_{i.}\right)^2}$$

where, Z_{ij} can have one of the following three definitions:

1. $Z_{ij} = \left|Y_{ij} - \overline{Y}_{i.}\right|$

   where, $\overline{Y}_{i.}$ is the mean of the i^{th} subgroup.

2. $Z_{ij} = \left|Y_{ij} - \tilde{Y}_{i.}\right|$

   where, $\tilde{Y}_{i.}$ is the median of the i^{th} subgroup.

3. $Z_{ij} = \left|Y_{ij} - \overline{Y}'_{i.}\right|$

where, $\overline{Y}'_{i.}$ is the 10% trimmed mean of the i^{th} subgroup. $\overline{Z}_{i.}$ are the group means of the Z_{ij} and $\overline{Z}_{..}$ is the overall mean of the Z_{ij}.

The three choices for defining Z_{ij} determine the robustness and power of Levene's test. By robustness, we mean the ability of the test to not falsely detect unequal variances when the underlying data are not normally distributed and the variables are in fact equal. By power, we mean the ability of the test to detect unequal variances when the variances are in fact unequal.

Using the trimmed mean which is performed best when the underlying data followed a Cauchy distribution (i.e., heavy-tailed) and the median which is performed best when the underlying data followed a χ^2_4 (i.e., skewed) distribution. Using the mean provided the best power for symmetric or moderate-tailed distributions.

Although the optimal choice depends on the underlying distribution, the definition based on the median is recommended as the choice that provides good robustness against many types of non-normal data while retaining good power. If you have knowledge of the underlying distribution of the data, this may indicate using one of the other choices.

The Levene's test rejects the hypothesis at level of significance α that the variances are equal if

$$W > F_{(\alpha, k-1, N-k)}$$

where, $F_{(\alpha, k-1, N-k)}$ is the upper critical value of the F distribution with $(k-1)$ and $(N-k)$ degrees of freedom at a significance level of α.

5.4.4 Assumptions of Normality

y_{ij} are jointly distributed in multivariate normal distribution. It is not possible to conduct exact test of significance based on the y_{ij} alone e.g. test of significance on Snedecor's F-distribution. Fortunately, normality, in addition to assumption 1 to 3 is sufficient for exact test of significance.

So, when assumptions 1 to 4 are all satisfied, then all of the usual analysis of variance procedure for estimating and testing to determine whether to infer the existence of, fixed linear relations e.g. non-zero difference among population means are strictly valid. In particular, an unbiased estimator of any given linear function of the parameters μ_{ij} is provided by the identical linear function of the observations y_{ij}, and an unbiased estimate of its variance can be derived from the residual or error mean square and exact confidence limits for the value of the given linear function of the parameters can be deduced with the aid to Student's t-distribution.

An easy method of checking the assumption of a single normal distribution is to construct a histogram of the data. The data should be examined for departures from normality before the tests are applied. However, the tests are robust to small departures from normality i.e. they work fairly well as long as the curve of the data is bell shaped and the tails are not heavy. Another method for testing the normality assumption is the normal probability plot. Goodness of fit for χ^2-test may be applied for testing normality. This assumption can be relaxed when the sample size is very large. If the assumptions do not hold, then a transformation of the y_{ij} into another scale will often allow ANOVA to be carried out. For more detailed use of transformations you can use Snedecor and Cochran (1980).

In many applications and applied microbiology in which bacterial number are being estimated, the assumption may not hold. The case when the sample sizes are small and in whole numbers with many zeros are unlikely to be normally distributed or wide range of observations may be present leading to heterogeneous variances.

Let us answer some activity questions.

E 2) Describe the assumptions of randomness in ANOVA.

E 3) Describe the assumptions of additivity in ANOVA.

E 4) Describe the assumptions of normality in ANOVA.

5.5 LINEAR MODELS USED IN ANALYSIS OF VARIANCE

Let $y_1, y_2, \ldots, y_n$ be the n observations of dependent variable/response variable. By using the additively assumptions of ANOVA, we shall assume that observed value y_{ij} be composed of two factors

$$y_{ij} = \mu_i + e_{ij}$$

where, μ_i is the true value and e_{ij} is the error. The true value μ_i is that part which is due to assignable causes and the portion that remains is the error, which is due to various chance causes. The true value μ_i is again assumed to be a linear function of k unknowns $T_1, T_2, \ldots, T_k$ called effects, i.e.

$$\mu_i = a_{i1}T_1 + a_{i2}T_2 + \ldots + a_{ik}T_k$$

where, a_{ij} are known and each being usually taken to be 0 or 1. This set up, which is fundamental to analysis of variance, is called linear model. So,

$$y_{ij} = a_{i1} T_1 + a_{i2} T_2 + \ldots + a_{ik} T_k + e_{ij}$$

5.5.1 Fixed Effect Model (Model 1)

A model in which all the effects T_i's are unknown constants, which we call parameters, is known as fixed effect model or model 1 or linear hypothesis model. It is often the case that one of the T_j's is constant and $a_{ij} = 1$ for that j and i. Such a T_j is called general mean or additive constant.

The fixed effect model of ANOVA applies to situations in which the experimenter applies one or more treatments (levels of a factor) to the subject/experimental units of the experiment to see if response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as whole.

5.5.2 Random Effects Models (Model 2) or Variance Components Model

A model in which all the effects T_i's are random variable except possibly the additive constant is called random effect model or model 2 or variance component model.

In Statistics, random effect model also called variance components model, is a kind of hierarchical linear model. It assumes that the data set /observations being analysed consist a hierarchy of different population whose differences relate to that hierarchy.

Random effects models are used when the treatments (levels of a factor) are not fixed. This occurs when the various factor levels are sampled from a larger population. Because the levels themselves are random variables, some assumptions and the method of contrasting, the treatments differ from ANOVA (model 1).

Suppose we are interested in knowing whether all the factors level effect (class effects) are equal or not. Now due to consideration of time, cost, or space, it is not possible to include in our experiment all the available factor levels (classes) effects. We can include only a sample of these factor levels and we want to infer about all the factor levels. Whether included in the experiment or not, form the result of the classes included in the experiment. In the random model, we shall consider balanced cases. Balanced cases are those cases, in which observations under different factor levels are the same. In higher order classification, if number of observations in cell is equal then it is called balanced.

In random effect models, main interest lies in estimation of variance components while in fixed effect models interest lies in estimation and testing the treatment differences. The difference of fixed effect model and random effect model can be seen from the following example:

Suppose a certain drug is thought to have an effect on the ability to perform mental arithmetic. The quantity of drug used may vary from 0 to 100 milligrams.

One possible experiment would be to test all possible levels of this drug on groups subjects and then use of analysis of variance to detect differences.

Because limited funds, only six levels of the drug may be tested. Then there will be two situations of implementation and analysis this experiment:

1. One possibility would be to systematical choosen a set of levels covering the range of doses. For example, one might choose 0, 20, 40, 60, 80 and 100. The differences in mental arithmetic scores could then be analysed using fixed effect model because doses have been fixed.
2. Another possibility is to choose the six dose levels randomly from the set of numbers 1 to 100. If this experiment were repeated then different levels might be chosen. The i^{th} dose levels changes and because of this the effect of i^{th} treatment is a random variable.

For instance, when an experimenter selects two or more treatments or two or more varieties, for testing, he rarely, if ever, draws them at random from a population of possible treatment or varieties he select those that he believes are most promising. In that situation, fixed effect model (model 1) is generally appropriate. On the other hand, when an experimenter selects a sample of varieties or treatments from a group of population for a study of the effects of various treatments or varieties, he can ensure that they are a random sample from the population of varieties or treatments by introducing randomisation into the sampling procedure for example, by using random number table. In this situation a random effect model (model 2) would clearly be appropriate.

Example of Random effect model

Suppose m large elementary schools are chosen randomly from among thousands in large country. Suppose also that n pupils of the same age are chosen randomly at each selected school. Their scores in a standard aptitude test are as curtained. Let y_{ij} be the score of the j^{th} pupil in the i^{th} school, then the following model is suggested:

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

where, μ is the average test score for the entire population. In this model α_i is the school specific random effect. It measures the difference between the average score at school i and the average score in the entire country and it is random because the school has been randomly selected from a larger populations of schools. The term, e_{ij}, is the individual specific error. That is, it is the deviation of the j^{th} pupil's score from the average for the i^{th} school. Again, this is regarded as random because of the random selection of pupils within the school, even though it is fixed quantity for any given pupil.

5.5.3 Mixed Effect Model

A model in which at least one T_i's is a random variable and at least one T_i' is constant (non-nagative constant) is called a mixed model.

Remark:

In general random effect is efficient, and should be used (over fixed effect) if the assumptions underlying it are believed to be satisfied. For random effect to work in the school example, it is necessary that the school specific effects be orthogonal to the other covariates of the model. This can be tested by running random effects, then fixed effects and doing a Housman specification to be seen test. If the test rejects, then random effects is blazed and fixed effects is the correct estimation procedure.

E 5) Describe the fixed effect model in ANOVA.

E 6) Describe the random effect model in ANOVA.

5.6 USES OF ANOVA

The following are some of the uses of ANOVA:

1. **To Test the Homogeneity of Several Means (k groups) or**

   $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

   If H_0 is rejected then we can say that there is a significant difference between these k groups or there is a significant effect of these k independent variables.

2. **To Test the Relationship between Two Variables**

   This test provides evidence that dependent variable Y_{ij} and independent variable X_{ij} are related in their movements. If Y_{ij} do not relate with X_{ij} then we expect $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$, which is the null hypothesis for testing the absence of relationship.

3. **Test for Linearity of Regression**

   After the relationship is established, the next step will be to find the appropriate regression function. At the first stage we try to find out whether the linear regression fit the observed data. So, the null hypothesis is now

   $H_0: \mu_i = \alpha + \beta X_i$

   with the sample model $Y_{ij} = \mu_i + e_{ij}$, when α and β are the parameters.

4. **Test for Polynomial Regression**

   The test procedure for testing the null hypothesis

   $H_0: \mu_i = \alpha + \beta_1 X_i + \beta_2 X_i^{\,2} + \ldots + \beta_k X_i^{\,k}$

   That the relationship between X and Y can be explained by a polynomial of degree k.

5. **Some Other Uses of ANOVA**

   - Test of homogeneity of a group of regression coefficients.
   - Test for equality of regression equations from p groups.
   - Test for multiple linear regression model.

E7) Explain briefly the uses of analysis of variance.

5.7 SUMMARY

In this unit, we have discussed:

1. The basic ideas and various types of analysis of variance;
2. Basic assumptions in analysis of variance;
3. Basic terminologies and notations in analysis of variance;
4. Various types of linear models in analysis of variance; and
5. Uses of ANOVA.

5.8 SOLUTIONS/ANSWERS

E 1) As same as Section 5.2 without Sub-sections 5.2.1 and 5.2.2.

E 2) As same as Sub-section 5.4.1.

E 3) As same as Sub-section 5.4.2.

E 4) As same as Sub-section 5.4.4.

E 5) As same as Sub-section 5.5.1.

E 6) As same as Sub-section 5.5.2.

E 7) As same as Section 5.6

UNIT 6 ONE-WAY ANALYSIS OF VARIANCE

Structure

6.1 Introduction
Objectives
6.2 One-way Analysis of Variance Model
6.3 Basic Assumptions of One-way Analysis of Variance
6.4 Estimation of Parameters
6.5 Test of Hypothesis
6.6 Degrees of Freedom of Various Sum of Squares
6.7 Expectations of Various Sum of Squares
Expectation of Treatment Sum of Squares
Expectation of Sum of Squares due to Error
6.8 ANOVA Table for One-way Classification
6.9 Summary
6.10 Solutions/Answers

6.1 INTRODUCTION

The analysis of variance is one of the powerful techniques of statistical analysis. Analysis of variance is used for testing of equality of means of several populations. It tests the variability of the means of the several populations. In the previous unit, we have discussed about the fundamental terms which are used in the analysis of variance. In that unit, we have also discussed the basic assumptions and models of analysis of variance.

As we have stated that the analysis of variance technique can be divided into two categories (i) parametric ANOVA and (ii) Non-parametric ANOVA. The parametric ANOVA can also be classified as one-way ANOVA if only one response variable is considered and MANOVA if two or more response variables are considered.

In this unit, we shall discuss the one-way analysis of variance. One-way analysis of variance is a technique where only one independent variable at different levels is considered which affects the response variable.

In this unit, the one-way analysis of variance model is discussed in Section 6.2. The basic assumptions under one-way analysis of variance are described in Section 6.3 whereas the estimates of each level mean of a factor are derived in Section 6.4. Test of hypothesis method is explained in Section 6.5 and the degrees of freedom for various sum of squares are determined in Section 6.6. The expectations of various sum of squares are derived in Section 6.7 whereas the analysis of variance table for one-way classification is described in Section 6.8.

Objectives

After studying this unit, you would be able to

- describe the one-way analysis of variance model;

- describe the basic assumptions under one-way analysis of variance;
- estimate of each level mean of a factor;
- determine the degrees of freedom for various sum of squares;
- derive the expectations of various sum of squares;
- construct the ANOVA table;
- test the hypothesis under one-way analysis of variance;
- identify differences in population means of k level of a factor; and
- determine which population means are different using multiple comparison methods.

6.2 ONE -WAY ANALYSIS OF VARIANCE MODEL

One-factor analysis of variance or one-way analysis of variance is a special case of ANOVA, for one factor of variable of interest and a generalization of the two sample t-test. The two sample t-test is used to decide whether two groups (two levels) of a factor have the same mean. One-way analysis of variance generalizes this to k levels (greater than two) of a factor.

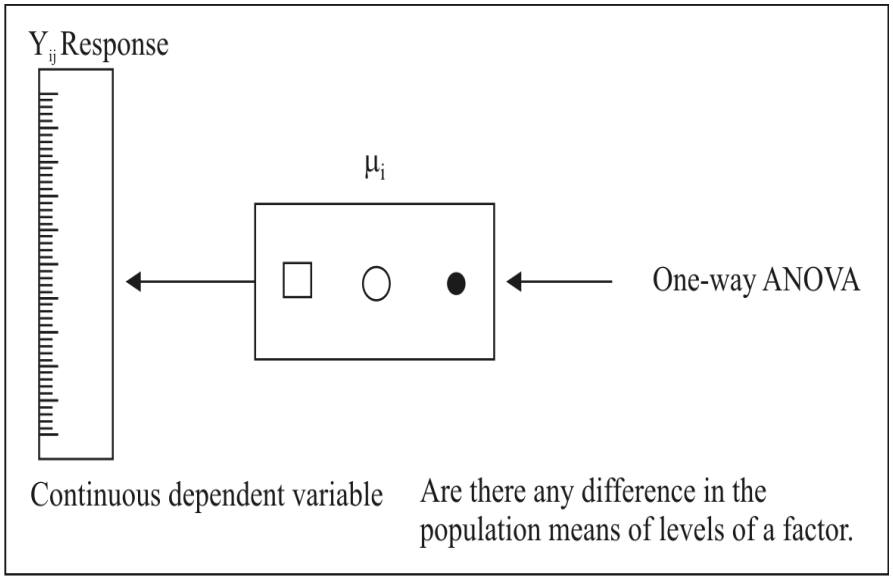
In the following, subscript i refers to the i^{th} level of the factor and subscript j refers to the j^{th} observation within a level of factor. For example y_{23} refers to third observation of the second level of a factor.

The observations on different levels of a factor can be exhibited below:

Level of a factor	**Observations**				**Totals**	**Means**
1	y_{11}	y_{12}	.....	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	.....	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
.	.	.	…….	.	.	.
.	.	.	…….	.	.	.
.	.	.	…….	.	.	.
i	y_{i1}	y_{i2}	…….	y_{in}	$y_{i.}$	$\bar{y}_{i.}$
.	.	.	…….	.	.	.
.	.	.	…….	.	.	.
.	.	.	…….	.	.	.
k	y_{k1}	y_{k2}	.....	y_{kn}	$y_{k.}$	$\bar{y}_{k.}$

The linear mathematical model for one-way classified data can be written as

$$y_{ij} = \mu_i + e_{ij} \qquad \text{where, } i = 1, 2, \ldots, k \ \& \ j = 1, 2, \ldots, n$$



Total observations are $n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = N$

Here, y_{ij} is continuous dependent or response variable, where μ_i is discrete independent variable, also called a explanatory variable.

This model decomposes the responses into a mean for each level of a factor and error term i.e.

Response = A mean for each level of a factor + Error term

The analysis of variance provides estimates for each level mean. These estimated level means are the predicted values of the model and the difference between the response variable and the estimated/predicted level means are the residuals.

That is

$$y_{ij} = \mu_i + e_{ij}$$

$$e_{ij} = y_{ij} - \mu_i$$

The above model can be written as $y_{ij} = \mu + (\mu_i - \mu) + e_{ij}$

or $y_{ij} = \mu + \alpha_i + e_{ij}, \quad \forall\ i = 1, 2, ..., k \text{ and } j = 1, 2, ..., n_i$

where, $\alpha_i = \mu_i - \mu$

A general mean effect given by

$$\mu = \frac{1}{N}\sum_{i=1}^{k} n_i \mu_i$$

This model decomposes the response into an over all (grand) mean, the effect of the i^{th} factor level α_i and error term $e_{ij.}$ The analysis of variance provides estimates of the grand mean μ and the effect of the i^{th} factor level α_i. The predicted values and the responses of the model are

$$y_{ij} = \mu + \alpha_i + e_{ij} \qquad \ldots (1)$$

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

α_i is the effect of the i^{th} level of the factor and given by

$$\alpha_i = \mu_i - \mu \qquad \forall \quad i = 1, 2, \ldots, k. \qquad \ldots (1a)$$

i.e. if the effect of i^{th} level of a factor increases or decreases in the yield/response variable by an amount α_i, then

$$\sum n_i \alpha_i = \sum n_i (\mu_i - \mu)$$

$$\sum n_i \alpha_i = \sum n_i \mu_i - \mu \sum n_i$$

$$= N\mu - N\mu$$

$$= 0$$

Under 6^{th} assumption, given in Section 6.3, the model becomes

$$E(y_{ij}) = \mu_i, \qquad \forall \quad i = 1, 2, \ldots, k \ \& \ j = 1, 2, \ldots, n_i.$$

or $$E(y_{ij}) = \mu + \alpha_i, \qquad \forall \quad i = 1, 2, \ldots, k \ \& \ j = 1, 2, \ldots, n_i.$$

6.3 BASIC ASSUMPTIONS OF ONE-WAY ANALYSIS OF VARIANCE

The following are the basic assumption of one-way ANOVA:

1. Dependent variable measured on interval scale;
2. k sample are independently and randomly drawn from the population;
3. Population can be reasonably to have a normal distribution;
4. k samples have approximately equal variance;
5. Various effects are additive in nature; and
6. e_{ij} are independently identically distributed normal with mean zero and variance σ_e^2.

Now, when we discuss the step by step computation procedure for one-way analysis of variance for k independent sample, the first step of the procedure is to make the null and alternative hypothesis.

We want to test the equality of the population means, i.e. homogeneity of effect of different levels of a factor. Hence, the null hypothesis is given by

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$

Against the alternative hypothesis

$$H_1: \mu_1 \neq \mu_2 \neq \ldots \neq \mu_k \text{ (or some } \mu_i\text{'s are not equal)}$$

which, reduces to

$$H_0: \alpha_1 = \alpha_2 \ldots . = \alpha_k = 0$$

Against the alternative hypothesis

$$H_1: \alpha_1 \neq \alpha_2 \ldots . \neq \alpha_k \neq 0 \text{ (or at least some } \alpha_i\text{'s are not zero)}$$

6.4 ESTIMATION OF PARAMETERS

The parameters μ and $\alpha_1, \alpha_2, ..., \alpha_k$ are estimated by the principle of least square on minimizing the error (residual) sum of squares. The residual sum of squares can be obtained as

$$e_{ij} = y_{ij} - \mu - \alpha_i \qquad \forall \quad i = 1, 2, \ldots, k \;\&\; j = 1, 2, \ldots, n_i.$$

$$e_{ij}^2 = (y_{ij} - \mu - \alpha_i)^2$$

$$E = \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right)^2$$

By this residual sum of squares E, we partially differentiate it with respect to μ and partially differentiating with respect to $\alpha_1, \alpha_2, ..., \alpha_k$ and then equating these equations to 0, we get

$$\frac{\partial E}{\partial \mu} = -2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \mu - \alpha_i\right) = 0$$

or $$\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} - N\mu - \sum_{i=1}^{k} n_i\alpha_i = 0\,, \quad \text{because} \quad kn_i = N$$

or $$\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} - N\mu = 0, \quad \text{because} \sum_{i=1}^{k} n_i\alpha_i = 0$$

or $$N\mu = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$$

Therefore, $$\mu = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \bar{y}_{..}$$

Similarly,

$$\frac{\partial E}{\partial \alpha_1} = -2\sum_{j=1}^{n_1}\left(y_{1j} - \mu - \alpha_1\right) = 0$$

$$\sum_{j=1}^{n_1} y_{1j} - n_1\mu - n_1\alpha_1 = 0$$

Or $$n_1\alpha_1 = \sum_{j=1}^{n_1} y_{1j} - n_1\mu$$

$$\alpha_1 = \frac{\sum_{j=1}^{n_1} y_{1j}}{n_1} - \hat{\mu}$$

Therefore, $\alpha_1 = \bar{y}_{1.} - \bar{y}_{..}$

Similarly,

$$\alpha_2 = \bar{y}_{2.} - \bar{y}_{..}$$

Or in general $\alpha_i = \bar{y}_{i.} - \bar{y}_{..} \qquad \forall \; i = 1, 2, \ldots, k.$

6.5 TEST OF HYPOTHESIS

Small differences between sample means are usually present. The objective is to determine whether these differences are significant or in other words, are the difference more than what might be expected to occur by chance? If the differences are more than what might be expected to occur by chance, you have sufficient evidence to conclude that there are differences between the population means of different levels of a factor.
The hypothesis is :

H_0: Population means of k levels of a factor are equal.

H_1: At least one population mean of a level of a factor is different from the population means of other levels of the factor.

or $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

$H_1: \mu_1 \neq \mu_2 = \ldots \neq \mu_k$

Now, substituting these estimated values in the model given in equation (1), the model becomes

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + e_{ij}$$

and then $$e_{ij} = (y_{ij} - \bar{y}_{..}) - (\bar{y}_{i.} - \bar{y}_{..})$$

$$= y_{ij} - \bar{y}_{i.}$$

Now substituting these values in equation (1) we get

So, $$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$$

Transporting by $\bar{y}_{..}$ to the left and squaring both sides and taking sum over i and j.

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})\right]^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..})$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 + 2\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})$$

But $\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.}) = 0$, since the sum of the deviations of the observations from their mean is zero.

Therefore,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k}n_i(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$$

where,

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$ is known as total sum of squares (TSS), $\sum_{i=1}^{k}n_i(\bar{y}_{i.} - \bar{y}_{..})^2$ is called between sum of squares or treatment sum of squares or sum of squares

due to different levels of a factor (SST) and $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2$ is called within sum of squares or residual sum of squares or error sum of squares (SSE).

So, TSS = SST + SSE

6.6 DEGREE OF FREEDOM OF VARIOUS SUM OF SQUARES

The Total Sum of Squares (TSS) which is computed from the N quantities of the form (y_{ij} - $\bar{y}$ ..) will carry (N−1) degrees of freedom. One degree of freedom lost because of linear constraints

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)=0$$

Similarly, the Treatment Sum of Squares (SST) = $\sum_{i=1}^{k}\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2$ will have (k−1) degrees of freedom.

Since $\sum_{i=1}^{k} n_i\left(\bar{y}_{i.}-\bar{y}_{..}\right)$ = 0 and the Error Sum of Squares i.e

$$SSE=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2$$

will have (N − k) degrees of freedom, since it is based upon N observations which are subject to k linear constraints

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right) = 0 \text{ for } i = 1, 2, ..., k$$

Hence, degrees of freedom of various sum of squares also additive same as various sum of squares (TSS = SST + SSE), i.e. degrees of freedom

$$(N-1) = (k-1) + (N-k)$$

Mean Sum of Squares

The sum of squares divided by its degrees of freedom is called Mean Sum of Squares (MSS). Therefore,

MSS due to treatment (MSST) = SST/df = SST/ (k−1).

MSS due to error (MSSE) = SSE/df = SSE/ (N−k).

6.7 EXPECTATIONS OF VARIOUS SUM OF SQUARES

For obtaining appropriate test statistics for testing H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k$, we have to find out expectations of various sum of squares. Our linear model is

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

We have

$$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(\mu + \alpha_i + e_{ij}\right)$$

$$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i}\mu + \frac{1}{n_i}\sum_{j=1}^{n_i}\alpha_i + \frac{1}{n_i}\sum_{j=1}^{n_i} e_{ij}$$

$$\bar{y}_{i.} = \mu + \alpha_i + \bar{e}_{i.} \qquad \ldots (2)$$

Now,

$$\bar{y}_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mu + \alpha_i + e_{ij}\right)$$

$$\bar{y}_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mu + \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\alpha_i + \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij} \quad \text{because} \sum_{i=1}^{k}\alpha_i = 0$$

$$\bar{y}_{..} = \mu + \bar{e}.. \qquad \ldots (3)$$

6.7.1 Expectation of Treatment Sum of Squares

$$E(SST) = E\left[\sum_{i=1}^{k} n_i\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2\right] \qquad \ldots (4)$$

Substituting the value of $\bar{y}_{i.}$ and $\bar{y}..$ from equations (2) & (3), we get

$$E\ (SST) = E\left[\sum_{i=1}^{k} n_i\left\{\left(\mu + \alpha_i + \bar{e}_{i.}\right) - \left(\mu + \bar{e}_{..}\right)\right\}^2\right]$$

$$= E\left[\sum_{i=1}^{k} n_i\left(\alpha_i + \bar{e}_{i.} - \bar{e}_{..}\right)^2\right]$$

$$= E\left[\sum_{i=1}^{k} n_i\left[\alpha_i^2 + \left(\bar{e}_{i.} - \bar{e}_{..}\right)^2 + 2\alpha_i\left(\bar{e}_{i.} - \bar{e}_{..}\right)\right]\right]$$

$$= \sum_{i=1}^{k} E\left(n_i\alpha_i^2\right) + \sum_{i=1}^{k} E\left(n_i\left(\bar{e}_{i.} - \bar{e}_{..}\right)^2\right) + 2\sum_{i=1}^{k} E\left(n_i\alpha_i\left(\bar{e}_{i.} - \bar{e}_{..}\right)\right)$$

$$= \sum_{i=1}^{k} n_i\alpha_i^2 + E\sum_{i=1}^{k} n_i\left(\bar{e}_{i.}^2 + \bar{e}_{..}^2 - 2\bar{e}_{i.}\bar{e}_{..}\right) + 2\sum_{i=1}^{k} n_i\alpha_i E\left(\bar{e}_{i.} - \bar{e}_{..}\right)$$

$$= \sum_{i=1}^{k} n_i\alpha_i^2 + E\sum_{i=1}^{k} n_i\left(\bar{e}_{i.}^2 + \bar{e}_{..}^2 - 2\bar{e}_{i.}\bar{e}_{..}\right) + 0 \quad \left[\text{because } E\left(\bar{e}_{i.} - \bar{e}_{..}\right) = 0\right]$$

$$= \sum_{i=1}^{k} n_i\alpha_i^2 + E\left[\sum_{i=1}^{k} n_i\bar{e}_{i.}^2 + \sum_{i=1}^{k} n_i\bar{e}_{..}^2 - 2\sum_{i=1}^{k} n_i\bar{e}_{i.}\bar{e}_{..}\right]$$

$$= \sum_{i=1}^{k} n_i\alpha_i^2 + \sum_{i=1}^{k} n_i E\left(\bar{e}_{i.}^2\right) + \sum_{i=1}^{k} n_i E\left(\bar{e}_{..}^2\right) - 2E\left[\sum_{i=1}^{k} n_i\bar{e}_{i.}\bar{e}_{..}\right]$$

$$= \sum_{i=1}^{k} n_i\alpha_i^2 + \sum_{i=1}^{k} n_i E\left(\bar{e}_{i.}^2\right) + NE\left(\bar{e}_{..}^2\right) - 2E\left[\bar{e}_{..}\sum_{i=1}^{k} n_i\bar{e}_{i.}\right]$$

$$= \sum_{i=1}^{k} n_i \alpha_i^2 + \sum_{i=1}^{k} n_i E\left(\bar{e}_{i.}^2\right) + N E\left(\bar{e}_{..}^2\right) - 2E\left(\bar{e}_{..} N\bar{e}_{..}\right) \quad \left(\text{because } \bar{e}_{..} = \frac{1}{N}\sum_{i=1}^{k} n_i \bar{e}_{i.}\right)$$

$$= \sum_{i=1}^{k} n_i \alpha_i^2 + \sum_{i=1}^{k} n_i E\left(\bar{e}_{i.}^2\right) - NE\left(\bar{e}_{..}^2\right) \quad \ldots (5)$$

Since, $e_{ij} \sim$ iid $N(0, \sigma_e^2)$ & $E(e_{ij}) = 0$

$V(e_{ij}) = E(e_{ij}^2) - (E(e_{ij}))^2$

$\sigma_e^2 = E(e_{ij}^2)$ … (6)

$E(\bar{e}_{i.}) = 0$ and $E(\bar{e}_{..}) = 0$

$V(\bar{e}_{i.}) = E(\bar{e}_{i.}^2) - (E(\bar{e}_{i.}))^2 = \sigma_e^2/n_i$

or $E(\bar{e}_{i.}^2) = \sigma_e^2/n_i$ … (7)

$V(\bar{e}_{..}) = E(\bar{e}_{..}^2) - (E(\bar{e}_{..}))^2 = \sigma_e^2/N$

or $E(\bar{e}_{..}^2) = \sigma_e^2/N$ … (8)

By Substituting values from equations (6), (7) and (8) in equation (5), we get

$$E(SST) = \sum_{i=1}^{k} n_i \alpha_i^2 + \sum_{i=1}^{k} n_i \frac{\sigma_e^2}{n_i} - N\frac{\sigma_e^2}{N}$$

$$= \sum_{i=1}^{k} n_i \alpha_i^2 + k\sigma_e^2 - \sigma_e^2$$

$$= \sum_{i=1}^{k} n_i \alpha_i^2 + (k-1)\sigma_e^2$$

Dividing both side by (k - 1), we get

$$E\left(\frac{SST}{k-1}\right) = \frac{1}{(k-1)}\sum_{i=1}^{k} n_i \alpha_i^2 + \sigma_e^2$$

$$E(MSST) = \sigma_e^2 + \frac{1}{(k-1)}\sum_{i=1}^{k} n_i \alpha_i^2$$

The second term in above equation like as variance since $\alpha_i = \mu_i - \mu$.
Therefore, under H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k$, it is zero.

MSS due to treatment provides an unbiased estimate of σ_e^2 under H_0.

6.7.2 Expectation of Sum of Squares due to Error

$$E(SSE) = E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2\right]$$

Substituting the value of y_{ij} and $\bar{y}_{i.}$ from equations (1) & (2), we get

$$E(SSE) = E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{\left(\mu + \alpha_i + e_{ij}\right) - \left(\mu + \alpha_i + \bar{e}_{i.}\right)\right\}^2\right]$$

$$= E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(e_{ij} - \bar{e}_{i.}\right)^2\right]$$

$$= E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(e_{ij}^2 + \bar{e}_{i.}^2 - 2e_{ij}.\bar{e}_{i.}\right)\right]$$

$$= E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 + \sum_{i=1}^{k} n_i \bar{e}_{i.}^2 - 2\sum_{i=1}^{k} \bar{e}_{i.}\left(\sum_{j=1}^{n_i} e_{ij}\right)\right]$$

$$= E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 + \sum_{i=1}^{k} n_i \bar{e}_{i.}^2 - 2\sum_{i=1}^{k} n_i \bar{e}_{i.}^2\right]$$

$$= E\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 - \sum_{i=1}^{k} n_i \bar{e}_{i.}^2\right]$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} E\left(e_{ij}^2\right) - \sum_{i=1}^{k} n_i E\left(\bar{e}_{i.}^2\right)$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} \sigma_e^2 - \sum_{i=1}^{k} n_i \frac{\sigma_e^2}{n_i}$$

$$= N\,\sigma_e^2 - k\,\sigma_e^2$$

$$= (N - k)\,\sigma_e^2$$

$$E\,(SSE/\,N-k) = \sigma_e^2$$

$$E\,(MSSE) = \sigma_e^2$$

Therefore, the error mean squares always gives an unbiased estimate of σ_e^2.

Under H_0, E (MSST) = E (MSSE)

Otherwise, E (MSST) > E (MSSE)

Hence, the test statistics for testing H_0 is provided by the variance ratio or Snedecor's

F = MSST/MSSE with [(k−1), (N−k)] df.

Thus, if an observed value of F is greater than the tabulated value of F for {(k−1), (N−k)} df and specific level of significance (usually 5% or 1%), then H_0 is rejected otherwise, it may be accepted.

6.8 ANOVA TABLE FOR ONE-WAY CLASSIFICATION

The above analysis is presented in the following table:

ANOVA Table for One-way Classified Data

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance Ratio F
Treatments	k−1	$\sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2 = SST$	MSST = SST/(k−1)	F = MSST/MSSE With{(k−1), (N−k)}df
Error	N−k	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 = SSE$	MSSE = SSE/(N−k)	
Total	N−1	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = TSS$		

If the treatment show significance effect at α level of significance then we are interested to find out which pair/pairs of treatments differ significantly, say the hypothesis H_{01}: $\mu_i = \mu_j$ or the null hypothesis is rejected.
With the help of

$$t = \frac{\bar{y}_{i.} - \bar{y}_{j.}}{\sqrt{MSSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \qquad \text{with df } N - k = k(n-1)$$

If $n_1 = n_2 = \ldots = n_k = n$ then this reduces to the

$$t = \frac{\bar{y}_{i.} - \bar{y}_{j.}}{\sqrt{\frac{2MSSE}{n}}}$$

or $$|t| = \frac{|\bar{y}_{i.} - \bar{y}_{j.}|}{\sqrt{\frac{2MSSE}{n}}} \text{ with } k(n-1) \text{ df}$$

If $t > t_{\alpha/2}[k(n-1)]$, then H_{01} is rejected at the level α. That is to say, H_{01} is rejected at the level of significance α if

$$|\bar{y}_{i.} - \bar{y}_{j.}| \geq |t|_{\alpha/2[k(n-1)]} \times \sqrt{\frac{2MSSE}{n}}$$

Thus, to compare the factor level effect/group means two at a time, we have to calculate $$|t|_{\alpha/2} \text{ (with df } k(n-1)) \times \sqrt{\frac{2MSSE}{n}}$$

which is called the critical difference (CD) or the least significant difference (lsd) and if the difference between the observed class/groups /factor level effect means, i.e. $(\bar{y}_{i.} - \bar{y}_{j.})$, is greater than the CD, then H_{01}: $\mu_i = \mu_j$ is rejected at α level of significance, otherwise it is accepted. Here, $t_{\alpha/2}[k(n-1)]$ is the tabulated value of t-distribution with k(n−1) df at upper α/2 point.

Example 1: An investigator is interested to know the level of knowledge about the history of India of 4 different schools in a city. A test is given to 5, 6, 7, 6 students of 8th class of 4 schools. Their scores out of 10 is given below:

School I (S_1)	8	6	7	5	9		
School II (S_2)	6	4	6	5	6	7	
School III(S_3)	6	5	5	6	7	8	5
School IV(S_4)	5	6	6	7	6	7	

Solution: If $\mu_1, \mu_2, \mu_3, \mu_4$ denote the average score of students of 8th class of schools I, II, III, IV respectively. Then

Null Hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Alternative hypothesis H_1: Difference among $\mu_1, \mu_2, \mu_3, \mu_4$ are significant.

S.No.	S_1		S_2	S_3	S_4	S_1^2	S_2^2	S_3^2	S_4^2
1	8		6	6	5	64	36	36	25
2	6		4	5	6	36	16	25	36
3	7		6	5	6	49	36	25	36
4	5		5	6	7	25	25	36	49
5	9		6	7	6	81	36	49	36
6			7	8	7		49	64	49
7				5				25	
Total	35		34	42	37	255	198	260	231

Grand Total G = 35 + 34 + 42 + 37 = 148

Correction Factor (CF) = $\frac{G^2}{N} = \frac{148^2}{24}$, Since $N = n_1 + n_2 + n_3 + n_4$

$= 912.6667$

Raw Sum of Squares (RSS) = $\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 = 255 + 198 + 260 + 231 = 944$

Total Sum of Squares (TSS) = RSS – CF = 944 – 912.6667 = 31.3333

Sum of Squares due to Treatments (SST)

$$= \frac{T_{1.}^2}{n_1} + \frac{T_{2.}^2}{n_2} + \frac{T_{3.}^2}{n_3} + \frac{T_{4.}^2}{n_4} - CF$$

$$= \frac{35^2}{5} + \frac{34^2}{6} + \frac{42^2}{7} + \frac{37^2}{6} - 912.6667$$

$$= 245 + 192.6667 + 252 + 228.1667 - 912.6667$$

$$= 5.1667$$

Sum of Squares due to Errors (SSE) = TSS − SST

$= 31.3333 - 5.1667 = 26.1666$

Now, $MSST = \frac{SST}{k-1} = \frac{5.1667}{3} = 1.7222$

$MSSE = \frac{SSE}{N-k} = \frac{26.1667}{20} = 1.3083$

ANOVA Table

Sources of Variation	DF	SS	MSS	F
Between schools	3	5.1667	1.7222	$F = \frac{1.7222}{1.3083} = 1.3164$
Within schools	20	26.1666	1.3083	

Calculated F = 1.3164

Tabulated F at 5% level of significance with (3, 20) degree of freedom is 3.10.

Conclusion: Since Calculated F < Tabulated F, so we may accept H_0 and conclude that level of knowledge of schools I, II, III and IV do not differ significantly.

Example 2: If we have three fertilizers and we have to compare their efficacy, this could be done by a field experiment in which each fertilizer is applied to 10 plots, and then 30 plots are later harvested, with the crop field being calculated for each plot. The data were recorded in following table:

Fertilizer	**Yields (in tones) from the 10 plots allocated to that fertilizer**									
1	6.27	5.36	6.39	4.85	5.99	7.14	5.08	4.07	4.35	4.95
2	3.07	3.29	4.04	4.19	.41	0.75	04.87	3.94	6.49	3.15
3	4.04	3.79	4.56	4.55	4.53	3.53	3.71	7.00	4.61	4.55

Solution:

H_0: Mean effect of Ist fertilizer = Mean effect of the IInd fertilizer = Mean effect IIIrd fertilizer

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

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

Steps for calculating different sum of squares

Grant Total = Total of all observation = $\sum\sum y_{ij}$ = G = 139.20

Correction Factor (CF) = G^2/N = 139.20 ×139.20 /30 = 645.89

Raw Sum of Squares (RSS) = $\sum\sum y_{ij}^2$ = 6385.3249

Total Sum of Squares = RSS − CF =36.4449

Sum of Squares due to Treatments/Fertilizer (SST)

$$= \frac{y_{1.}^2}{10} + \frac{y_{2.}^2}{10} + \frac{y_{3.}^2}{10} - CF$$
$$= (54.5)^2/10 + (40)^2/10 + (44.9)^2/10 - CF$$
$$= 10.8227$$

Sum of Squares due to Error = TSS − SST = 36.4449 −10.8227

= 25.6222

Mean Sum of Squares due to Treatments/Fertilizers (MSST) = SST/df

= 10.8227/2 = 5.4114

Mean Sum of Squares due to Error (MSSE) = SSE/df

= 25.6221/27 = 0.9490

Variance ratio $F_{2,27}$ = MSST/MSSE = 5.414/0.9490 = 5.70

Tabulated $F_{2,27}$ = 3.35

Since calculated value of $F_{2,27}$ is greater than tabulated $F_{2,27}$ at 5% level of significance so we reject H_0. It means there is a significant difference among the effect of these three fertilizers.

Now, H_0 is rejected. So, we have to test which of the fertilizers most important among the pairwise comparison. So, pairwise comparison test will be applied.

To test $H_0: \mu_i = \mu_j$

against $H_1: \mu_i \neq \mu_j$

We have a test statistic

$$\left|\bar{y}_{i.} - \bar{y}_{j.}\right| \geq \left|t\right|_{\alpha/2}(N-k) \times \sqrt{\frac{2MSE}{10}}$$

or $$\left|\bar{y}_{1.} - \bar{y}_{2.}\right| = 5.45 - 4.0 = 1.45 \quad \ldots (9)$$

$$\left|\bar{y}_{2.} - \bar{y}_{3.}\right| = (4.00 - 4.49) = 0.49 \quad \ldots (10)$$

$$\left|\bar{y}_{1.} - \bar{y}_{3.}\right| = (5.45 - 4.49) = 0.96 \quad \ldots (11)$$

and

$$t_{\alpha/2} \text{ at } 27 \text{ df} \times \sqrt{\frac{2(0.9490)}{10}} = 2.05 \times \sqrt{\frac{1.8980}{10}}$$

$$= 2.05 \times 0.044 = 0.9020$$

Since, the calculated value of difference of mean effect 1 with 2 and 3 is greater than the value of the critical difference (0.9020). So, effect of fertilizer 1 has significant difference with effect of fertilizers 2 and 3. But there is no significant difference between the effects of fertilizer 2 and 3. Since mean value of fertilizer 1 $\left(\bar{y}_{1.} = 5.47\right)$ is greater than mean values of fertilizer 2 or 3 so we can say that fertilizer 1 can be preferred in comparison to fertilizer 2 and 3.

E 1) Three varieties A, B and C of wheat are shown in five plots each of the following fields per acre as obtained:

Plots	A	B	C
1	8	7	12
2	10	5	9
3	7	10	13
4	14	9	12
5	11	9	14

Set up a table of analysis of variance and find out whether there is significant difference between the fields of these varieties.

E 2) The following figures relate to production in kg. of three varieties P, Q, R of wheat shown in 12 plots

P	14	16	18		
Q	14	13	15	22	
R	18	16	19	15	20

Is there any significant difference in the production of these varieties?

E 3) In 25 plots four varieties ν_1, ν_2, ν_3, ν_4 of wheat are randomly put and their yield in kg are shown below.

ν_1 2000	ν_3 2270	ν_2 2230	ν_4 2270	ν_4 2180
ν_2 2160	ν_1 2100	ν_2 2050	ν_3 2300	ν_2 2280
ν_1 2200	ν_1 2300	ν_4 2040	ν_3 2420	ν_1 2240
ν_4 2370	ν_1 2250	ν_2 2040	ν_2 2360	ν_1 2460
ν_3 2210	ν_1 2340	ν_2 2190	ν_1 2150	ν_3 2020

Perform the ANOVA to test whether there is any significant difference between varieties of wheat.

6.9 SUMMARY

In this unit, we have discussed:

1. The one-way analysis of variance model;
2. The basic assumptions in one-way analysis of variance;
3. Estimation of parameters of one-way analysis of variance model;
4. Test of hypothesis for one-way classified ANOVA;
5. How to obtain the expectation of various sum of squares in one-way ANOVA; and
6. The construction of one-way ANOVA table.

6.10 SOLUTIONS /ANSWERS

E1) Null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ i.e. the mean fields of three varieties is the same,

Against the alternative hypothesis $H_1 : \mu_1 \neq \mu_2 \neq \mu_3$

The calculation is done on the basis of the given data and the results are as follows:

$G = \sum\sum y_{ij}$ = Sum of all observations

G = 8+10+7+14+11+7+5+10+9+9+12+9+13+12+14 = 150

N = Total number of observations = 15

Correction factor $(CF) = \frac{G^2}{N} = \frac{150 \times 150}{15} = 1500$

Raw Sum of Squares (RSS) $= \sum\sum y_{ij}^2$

$= 8^2 + 10^2 + 7^2 + 14^2 + 11^2 + 7^2 + 5^2 + 10^2$

$$+9^2+9^2+12^2+9^2+13^2+12^2+14^2$$

$$= 1600$$

Total Sum of Squares (TSS) = Raw Sum of Squares – CF

$$= 1600 - 1500 = 100$$

Sum of Squares due to Treatments (SST) = $\frac{T_A^2}{5}+\frac{T_B^2}{5}+\frac{T_C^2}{5}-CF$

$$= \frac{(50)^2}{5}+\frac{(40)^2}{5}+\frac{(60)^2}{5}-CF$$

$$= \frac{1}{5}[2500+1600+3600]-1500$$

$$= 1540 - 1500 = 40$$

Sum of Squares due to Error (SSE) = TSS - SST

$$= 100 - 40 = 60$$

Mean Sum of Squares due to Treatments (MSST) = $\frac{SST}{df} = 20$

Mean Sum of Squares due to Error (MSSE) = $\frac{MSSE}{df} = \frac{60}{12} = 5$

Therefore, $F = \frac{MSST}{MSSE} = \frac{20}{5} = 4$

ANOVA Table for One-way Classified Data

Sources of Variation (SV)	**Degrees of Freedom**	**Sum of Squares**	**Mean Sum of Squares (MSS)**	**F-Statistic or Variation Ratio**
Due to three varieties or due to treatments	2	40 (SST) or (SSB)	MSST = 20	$F_{2,12} = \frac{20}{5} = 4$
Due to error within groups	12	60 (SSE) or (SSW)	MSSE = 5	
Total	14	TSS = 100		

For $\nu_1 = 2, \quad \nu_2 = 12,$ the table value of F at 5% level of significance is 3.88 which can be seen from the statistical table. Since the calculated value is greater than the table value of F at 5% level of significance. So, we reject the null hypothesis and hence we conclude that the difference between the mean field of three varieties is significant.

Since the null hypothesis is rejected, then pairwise comparison test may be applied for testing the null hypothesis of equality of two population means. For this, critical difference (CD) will be calculated by using the formula

$$CD = \sqrt{\frac{2MSSE}{5}} \times t_{0.05} \text{ at error df}$$

$$CD = \sqrt{\frac{2\times 5}{5}} \times 2.571$$

$$= \sqrt{2} \times 2.571 = 1.41 \times 2.571$$

$$= 3.625$$

$$\left|\overline{T_1} - \overline{T_2}\right| = |10 - 8| = 2$$

$$\left|\overline{T_1} - \overline{T_3}\right| = |10 - 12| = 2$$

$$\left|\overline{T_2} - \overline{T_3}\right| = |8 - 12| = 4$$

Since $\left|\overline{T_1} - \overline{T_2}\right|$ and $\left|\overline{T_1} - \overline{T_3}\right|$ are less than CD. So we accept the null hypothesis which means that if we interested to take out of A and B varieties then we can take any of these two. Similarly, between A and C we can take any of varieties. But if we conclude to take out of B and C then we should prefer C because hypothesis of equality two mean is rejected and the mean value corresponding to C varieties is higher than B varieties.

E2) Null hypotheses H_0: $\mu_i = \mu_2 = \mu_3$, i.e. there is no difference in the production of these varieties P, Q and R against the alternative hypothesis $H_1 : \mu_i \neq \mu_2 \neq \mu_3$

$G = \sum y_{ij}$ = Grand total

= 14+16+18+14+13+15+22+18+16+19+19+20 = 204

N = Total number of observations = 12

Correction Factor (CF) = $\frac{G^2}{N} = \frac{204\times 204}{12} = 3468$

Raw Sum of Squares (RSS) = $\sum\sum y_{ij}^{\ 2}$

$$= 14^2+16^2+18^2+14^2+13^2+15^2+22^2$$
$$+18^2+16^2+19^2+19^2+20^2$$
$$= 3552$$

Total Sum of Squares (TSS) = RSS - CF

= 3552 - 3468 = 84

Sum of Squares due to Treatments (SST)

$$= \frac{T_P^2}{3} + \frac{T_Q^2}{4} + \frac{T_R^2}{5} - CF$$

$$= \frac{48\times 48}{3} + \frac{64\times 64}{4} + \frac{92\times 92}{5} - 3468$$

$$= 768+1024+1692.80 - 3468 = 16.8$$

Sum of Squares due to Error (SSE) = TSS - SST

= 84 - 16.8 = 67.2

Mean Sum of Squares due to Treatments (MSST) = $\frac{SST}{df} = \frac{16.8}{2} = 8.4$

Mean Sum of Squares due to Error (MSSE) = $\frac{SSE}{df} = \frac{67.2}{9} = 7.467$

Therefore,

$$F_{2,9} = \frac{MSST}{MSSE} = \frac{8.4}{7.467} = 1.125$$

ANOVA Table for One-way Classified Data

Sources of Variation (SV)	**Degrees of Freedom**	**Sum of Squares (SS)**	**Mean Sum of Squares (MSS)**	**F-statistic or Variation Ratio**
Between Varieties	2	16.8	8.4	$F_{2,9} = \frac{8.4}{7.46} = 1.12$
Due to Error	12	67.20	7.467	
Total	14	84		

For $\nu_1 = 2$ and $\nu_2 = 9$ the tabulated value of F at 5% level significance is 4.261. Since the calculated value F is less than the table value of F, we accept the null hypotheses and conclude that there is no any significance difference in their mean productivity of three varieties P, Q and R.

E3) We have

ν_1	ν_2	ν_3	ν_4
2000	2160	2210	2370
2200	2230	2270	2040
2100	2050	2300	2270
2300	2040	2420	2180
2250	2190	2020	
2340	2360		
2150	2280		
2240			
2460			

To simplify the calculations, we subtract some suitable number 2200 (say) from all observations and then dividing by 10 we have

S.No.	v_1	v_2	v_3	v_4	$(v_1)^2$	$(v_2)^2$	$(v_3)^2$	$(v_4)^2$
1	-20	-4	1	17	400	16	1	289
2	0	3	7	-16	00	9	49	256
3	-10	-15	10	7	100	225	100	49
4	10	-16	22	-2	100	256	484	4
5	5	-1	-18		25	1	324	
6	14	16			196	256		
7	-5	8			25	64		
8	4				16			
9	26				676			
Total	24	-9	22	6	1538	827	958	598

Null Hypothesis H_0: There is no significant difference in the effect of varieties.

Against the alternative hypothesis H: There is significant difference in the effect of varieties.

$$G = 24 + (-9) + 22 + 6 = 43$$

$$CF = \frac{G^2}{N} = \frac{(43)^2}{25} = 73.96$$

$$RSS = 1538 + 827 + 958 + 598 = 3921$$

$$TSS = RSS - CF = 3921 - 73.96 = 3847.04$$

$$SST = \frac{(24)^2}{9} + \frac{(-9)^2}{7} + \frac{(22)^2}{5} + \frac{(6)^2}{4} - 73.96$$

$$= 64 + 11.5714 + 96.8 + 9 - 73.96$$

$$= 107.4114$$

$$SSE = TSS - SST$$

$$= 3847.04 - 107.4114$$

$$= 3739.6286$$

$$MSST = \frac{SST}{df} = \frac{107.4114}{3} = 35.8038$$

$$MSSE = \frac{SSE}{df} = \frac{3739.6284}{21} = 178.0776$$

$$F = \frac{MSST}{MSSE} = \frac{35.8038}{178.0776} = 0.2011$$

ANOVA Table

Source of variation	SS	df	MSS	F
Between Varieties	107.4114	3	35.8038	F =0.2011
Due to errors	3739.6286	21	178.0776	
Total	3847.04	24		

Calculated F = 0.2011

Tabulated value of F at 5% level of significance with (3, 21) degrees of freedom is 3.07

Conclusion: Since calculated $|F|$ < Tabulated F, so we may accept H_0 and conclude that varieties ν_1, ν_2, ν_3, ν_4 of wheat are homogeneous.

UNIT 7 TWO-WAY ANALYSIS OF VARIANCE

Structure

7.1 Introduction
Objectives
7.2 Two-way ANOVA Model with One Observation per Cell
7.3 Assumptions of Two-way ANOVA
7.4 Estimation of Parameters in Two-way Classified Data
7.5 Test of Hypothesis in Two-way ANOVA
7.6 Degrees of Freedom of Sum of Squares
7.7 Expectations of Various Sum of Squares
Expectation of Sum of Squares due to Factor A
Expectation of Sum of Squares due to Factor B
Expectation of Sum of Squares due to Error
7.8 Summary
7.9 Solutions /Answers

7.1 INTRODUCTION

In Unit 6 of this block, we have discussed the one-way analysis of variance. In one-way analysis of variance, we have considered one independent variable at different levels which affects the response variable. Analysis of variance is a technique which split up the total variation of data which may be attributed to various “sources” or “causes” of variation. There may be variation between variables and also within different levels of variables. In this way, analysis of variance is used to test the homogeneity of several population means by comparing the variances between the sample and within the sample. In this unit, we will discuss the two-way analysis of variance technique. In two-way analysis of variance technique, we will consider two variables at different levels which affect the response variables.

In Section 7.2 the two-way ANOVA model is explained whereas the basic assumptions in two-way ANOVA are described in Section 7.3. The estimates of each level mean of each factor are found in Section 7.4. Test of hypothesis in two-way ANOVA is explained in Section 7.5 and the degrees of freedom of various sum of squares in two-way ANOVA are described in Section 7.6. Expectations of various sum of squares are derived in Section 7.7.

Objectives

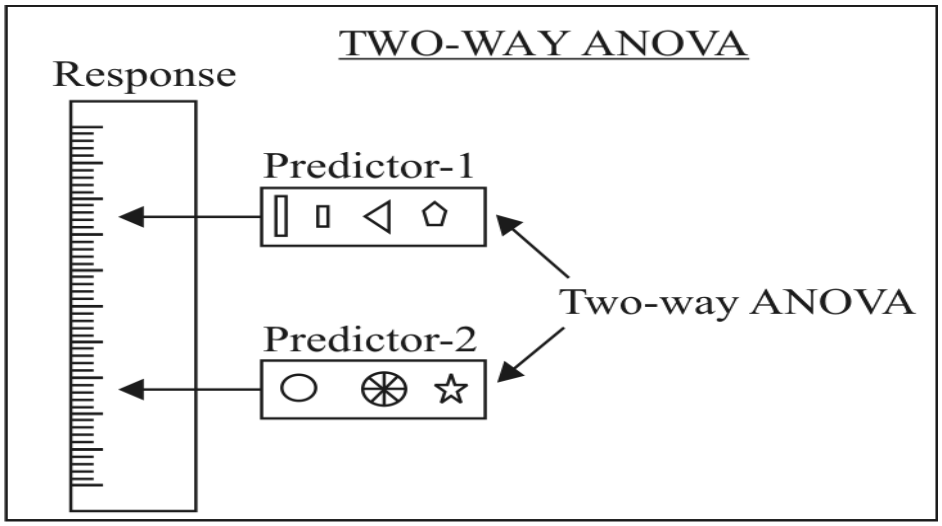
After reading this unit, you would be able to

- explain the two-way ANOVA Model;
- describe the basic assumptions in two-way ANOVA;
- find the estimate of each level mean of each factor;
- explain the degrees of freedom of various sum of squares in two-way ANOVA;

- derive the expectations of various sum of squares;
- construct the ANOVA Table for two-way classification and determine the critical differences; and
- perform the two-way ANOVA test.

7.2 TWO-WAY ANOVA MODEL WITH ONE OBSERVATION PER CELL

In the previous unit, we have considered the case where you had one categorized predictor/independent variable/explanatory at different levels. In this unit, consider a case with two categorical predictors. In general, some times you have more than one categorical predictor variables at a different levels and a continuous response variable then it is called two-way classification.



Two-way classified data can be treated in three ways

1. Analysis of two-way classified data with one observation per cell.
2. Analysis of two-way classified data with equal number of observations (m observation) per cell.
3. Analysis of two-way classified data with unequal number of observation per cell.

In this unit we shall discuss the first case i.e. Case I: Analysis of two-way classified data with one observation per cell (Fixed effect model) only.

We have an experiment in such a way as to study the effect of two factors in the same experiment. For each factor, there will be a number of classes/groups or levels. In the fixed effect model, there will be only fixed levels of the two factors. We shall first consider the case of one observation per cell. Let the factors be A and B and the respective levels be $A_1, A_2, \ldots, A_p$ and $B_1, B_2, \ldots, B_q$. Let y_{ij} be the observation/response/dependent variable under the i^{th} level of factor A and j^{th} level of factor B. The observations can be represented as follows:

Table of Two-Way Classified Data

A/B	B_1	B_2	…	B_j	…	B_q	Total	Mean
A_1	y_{11}	y_{12}	…	y_{1j}	…	y_{1q}	$y_{1.}$	$\bar{y}_{1.}$
A_2	y_{21}	y_{22}	…	y_{2j}	…	y_{2q}	$y_{2.}$	$\bar{y}_{2.}$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
A_i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{iq}	$y_{i.}$	$\bar{y}_{i.}$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
A_p	y_{p1}	y_{p2}	…	y_{pj}	…	y_{pq}	$y_{p.}$	$\bar{y}_{p.}$
Total	$y_{.1}$	$y_{.2}$	…	$y_{.j}$	…	$y_{.q}$	$y_{..} = G$	
Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	…	$\bar{y}_{.j}$	…	$\bar{y}_{.q}$		$\bar{y}_{..}$

Mathematical Model

Here, the mathematical model may be written as

$$y_{ij} = \mu_{ij} + e_{ij}$$

where, e_{ij} are independently normally distributed with common mean zero and variance σ_e^2. Corresponding to above table of observations following table is a table of expected values of observations:

Table of Expectations of Observations of y_{ij}

A/B	B_1	B_2	…	B_j	…	B_q	Mean	Difference
A_1	μ_{11}	μ_{12}	…	μ_{1j}	…	μ_{1q}	$\mu_{1.}$	$\mu_{1.} - \mu = \alpha_1$
A_2	μ_{21}	μ_{22}	…	μ_{2j}	…	μ_{2q}	$\mu_{2.}$	$\mu_{2.} - \mu = \alpha_2$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
A_i	μ_{i1}	μ_{i2}	…	μ_{ij}	…	μ_{iq}	$\mu_{i.}$	$\mu_{i} - \mu = \alpha_i$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
A_p	μ_{p1}	μ_{p2}	…	μ_{pj}	…	μ_{pq}	$\mu_{p.}$	$\mu_{p.} - \mu = \alpha_p$
Mean	$\mu_{.1}$	$\mu_{.2}$	…	$\mu_{.j}$	…	$\mu_{.q}$	$\mu_{..}$	
Difference	$\mu_{.1} - \mu = \beta_1$	$\mu_{.2} - \mu = \beta_2$	…	$\mu_{.j} - \mu = \beta_j$	…	$\mu_{.q} - \mu = \beta_q$		

Now μ_{ij} can be written as

$$\mu_{ij} = \mu + (\mu_{i.} - \mu) + (\mu_{.j} - \mu) + (\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu)$$

$$= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

or $$E(y_{ij}) = \mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

or $$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ij}$$

where μ is a constant general mean effect, present in all the observations, $\alpha_i = \mu_{i.} - \mu$ is an effect due to the i^{th} level of the factor A, which is common to all the observations belonging to this level of A , $\beta_j = \mu_{.j} - \mu$ is an effect due to j^{th} level of the factor B, which is common to all the observations belonging of this levels of B and $(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i.} + \mu_{.j} + \mu$ is called the interaction between the i^{th} level of A and j^{th} level of B. It is an effect peculiar to the combination $(A_i B_j)$. It is not present in the i^{th} level of A or in the j^{th} level of B if not taken together i.e.if the joint effect of A_i and B_j is different from the sum of the effects due to A_i and B_j taken individually, it means there is interaction and it is measured by $(\alpha\beta)_{ij}$.

In the case of two-way classified data with one observation per cell, the interaction $(\alpha\beta)$ cannot be estimated and, therefore, there is no interaction i.e. $(\alpha\beta)_{ij}$ for all i and j. So, the model becomes

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \qquad \ldots (1)$$

7.3 ASSUMPTIONS OF TWO-WAY ANOVA

For the validity of F-test, the following assumptions should be satisfied:

1. All the observations y_{ij} are independent;
2. Different effects (effects of levels of factor A, effects of levels of factor B and error effect) are additive in nature;
3. e_{ij} are independently and identically distributed normally with mean zero and variance σ_e^2 , i.e. $e_{ij} \approx iid\, N(0, \sigma_e^2)$; and
4. There is no interaction between levels of factor A and the levels of factor B.

7.4 ESTIMATION OF PARAMETERS IN TWO-WAY CLASSIFIED DATA

From the above model given in equation (1)

$$e_{ij} = y_{ij} - \mu - \alpha_i - \beta_j \qquad i = 1, 2, \ldots, p$$

Squaring both the sides $\qquad j = 1, 2, \ldots, q$

$$e_{ij}^{\ 2} = (y_{ij} - \mu - \alpha_i - \beta_j)^2$$

Summing over i & j both the sides, we get

$$\sum_{i=1}^{p}\sum_{j=1}^{q} e_{ij}^{\ 2} = \sum_{i=1}^{p}\sum_{j=1}^{q} (y_{ij} - \mu - \alpha_i - \beta_j)^2$$

Let us suppose that this can be written as

$$E = \sum_{i=1}^{p}\sum_{j=1}^{q}\left(y_{ij} - \mu - \alpha_i - \beta_j\right)^2$$

For obtaining the least square estimates, we minimize E. For minimizing E, differentiating it with respect to μ we get

$$\frac{\partial E}{\partial \mu} = -2\sum_{i=1}^{p}\sum_{j=1}^{q}\left(y_{ij} - \mu - \alpha_i - \beta_j\right)$$

and now equating this equal to zero, we have

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\left(y_{ij} - \mu - \alpha_i - \beta_j\right) = 0$$

or $$\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij} - pq\mu - q\sum_{i=1}^{p}\alpha_i - p\sum_{j=1}^{q}\beta_j = 0$$

It is clear from the above table that

$$\mu_{i.} = \mu - \alpha_i \quad \text{or} \quad \sum_{i=1}^{p}\left(\mu_{i.} - \mu\right) = \sum_{i=1}^{p}\alpha_i$$

or $$\sum_{i=1}^{p}\mu_{i.} - p\mu = \sum_{i=1}^{p}\alpha_i$$

or $$p\mu - p\mu = \sum_{i=1}^{p}\alpha_i$$

or $$\sum_{i=1}^{p}\alpha_i = 0$$

Similarly, $$\mu_{.j} - \mu = \beta_j \Rightarrow \sum_{j=1}^{q}\beta_j = 0$$

Using these relations and substituting in the equation

$$\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij} - pq\mu - q\sum_{i=1}^{p}\alpha_i - p\sum_{j=1}^{q}\beta_j = 0$$

$$\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij} = pq\hat{\mu}$$

or $$\hat{\mu} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij}}{pq} = \bar{y}_{..}$$

Similarly, differentiating E with respect to $\alpha_1, \alpha_2, ..., \alpha_p$, we get

for i=1, i.e. for estimating α_1

$$E = \sum_{j=1}^{q}(y_{1j} - \mu - \alpha_1 - \beta_j)^2$$

$$\frac{\partial E}{\partial \alpha_1} = -2\sum_{j=1}^{q}\left(y_{1j} - \mu - \alpha_1 - \beta_j\right)$$

Now equating this equation to zero, we get

$$\sum_{j=1}^{q}\left(y_{1j} - \mu - \alpha_1 - \beta_j\right) = 0$$

$$\sum_{j=1}^{q} y_{1j} - q\hat{\mu} - q\hat{\alpha}_1 - \sum_{j=1}^{q}\beta_j = 0$$

or $\sum_{j=1}^{q} y_{1j} - q\hat{\mu} = \hat{\alpha}_1$ because $\sum_{j=1}^{q} \beta_j = 0$

or, $\hat{\alpha}_1 = \frac{\sum_{j=1}^{q} y_{1j}}{q} - \hat{\mu} = (\bar{y}_{1.} - \bar{y}_{..})$

or in general $\hat{\alpha}_i = (\bar{y}_{i.} - \bar{y}_{..})$ for all i =1, 2, ..., p.

Similarly for obtaining the least square estimate for β_j we have to differentiate E w.r.t. β_j and equating to zero, we have

$$\frac{\partial E}{\partial \beta_j} = -2\sum_{i=1}^{p}(y_{ij} - \mu - \alpha_i - \beta_j) = 0$$

or $\sum_{i=1}^{p} y_{ij} - \sum_{i=1}^{p} \mu - \sum_{i=1}^{p} \alpha_i - \sum_{i=1}^{p} \beta_j = 0$

or $\sum_{i=1}^{p} y_{ij} - p\hat{\mu} - p\hat{\beta}_j = 0$ because $\sum_{i=1}^{p} \alpha_i = 0$

or $\hat{\beta}_j = \frac{1}{p}\sum_{i=1}^{p} y_{ij} - \hat{\mu}$

$\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$ for all j =1, 2, . . . , q

7.5 TEST OF HYPOTHESIS IN TWO-WAY ANOVA

It is clear that $\sum_{i=1}^{p} \alpha_i = 0$, $\sum_{j=1}^{q} \beta_j = 0$ and $\sum_{i=1}^{p}\sum_{j=1}^{q} (\alpha\beta)_{ij} = 0$

So, we have the model

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

The least square estimators, as already obtained in previous section by minimizing

$$E = \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \mu - \alpha_i - \beta_j)^2$$

and differentiating E with respect to μ, α_i, β_j , we get

$$\hat{\mu} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij}^2}{pq} = \bar{y}_{..}$$

$\hat{\alpha}_i = (\bar{y}_{i.} - \bar{y}_{..})$ for all i =1, 2 ,…, p.

$\hat{\beta}_j = (\bar{y}_{.j} - \bar{y}_{..})$ for all j =1, 2, … , q.

In this model each observation is the sum of four components and analysis of variance partitions $\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij}^2$ also in four components as follows:

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

or $$y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

Squaring both sides and summing over i and j

$$\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{.j} - \bar{y}_{..})^2$$

$$+ \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 + 2\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{.j} - \bar{y}_{..})$$

$$+ 2\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{i.} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

$$+ 2\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{.j} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

Therefore,

$$\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{..})^2 = q\sum_{i=1}^{p}(\bar{y}_{i.} - \bar{y}_{..})^2 + p\sum_{j=1}^{q}(\bar{y}_{.j} - \bar{y}_{..})^2$$

$$+ \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

Other terms are zero because, sum of the deviation of observations from their mean is always zero.

So, Total Sum of Squares = Sum of Squares due to Factor A (SSA)

+ Sum of Squares due to Factor B (SSB)

+ Sum of Squares due to Error (SSE)

or in short TSS = SSA + SSB + SSE

The corresponding partition on the total degrees of freedom is as follows:

df of TSS = df of SSA + df of SSB + df of SSE

$pq-1 = (p-1) + (q-1) + (p-1)(q-1)$

Dividing the sum of squares by their respective degrees of freedom (df), we get corresponding mean sum of squares.

In two-way classified data, we can test two hypotheses, one for levels of factor A that is equality of different levels of factor A

H_{0A}: $\alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$

H_{1A}: $\alpha_1 \neq \alpha_2 \neq \ldots \neq \alpha_p \neq 0$

and for equality of different levels of factor B

H_{0B}: $\beta_1 = \beta_2 = \ldots = \beta_q = 0$

H_{1B}: $\beta_1 \neq \beta_2 \neq \ldots \neq \beta_q \neq 0$

To derive the appropriate tests for these hypotheses we have obtained the expected values of various sum of squares as explained in one-way classified data in Section 7.7.

So,

$$E(MSSA) = \sigma_e^2 + q\sum_{i=1}^{p} \alpha_i^2 / (p-1)$$

$$E(MSSB) = \sigma_e^2 + p\sum_{j=1}^{q} \beta_j^2 / (q-1)$$

$$E(MSSE) = \sigma_e^2$$

If H_{0A}: $\alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$ is true, then E (MSSA) = E (MSSE) and hence

F = MSSA / MSSE will give the test to test H_{0A}.

So, a test for the hypothesis of equality of the effect of the different levels of factor A is provided by this F, which follows the Snedecor F- distribution with [(p - 1), (p - 1)(q - 1)] df.

Thus, the null hypothesis will be rejected at α level of significance if and only if F = MSSA/MSSE > F tabulated [(p - 1), (p - 1) (q - 1)] df at α level of significance.

Similarly if, H_{0B}: $\beta_1 = \beta_2 = \cdots = \beta_q = 0$ for the equality of the effects of the different levels of B is true then E(MSSB) = E(MSSE) and F = MSSB/MSSE will give the test to test H_{0B}. H_{0B} is rejected at the α level of significance if and only if F = MSSB/MSSE > F tabulated [(q - 1), (p - 1) (q - 1)]df at α level of significance. These calculations are shown in the following table.

ANOVA Table for Two-way Classified Data with One Observation per Cell

Source of Variation	DF	SS	MSS	F
Between the Levels of A	p - 1	$SSA = q\sum_{i=1}^{p}(\bar{y}_{i.} - \bar{y}_{..})^2$	MSSA = SSA/(p - 1)	$F_{(p-1),(p-1)(q-1)}$ = MSSA/MSSE
Between the Levels of B	q - 1	$SSB = p\sum_{j=1}^{q}(\bar{y}_{.j} - \bar{y}_{..})^2$	MSSB =SSA/(q - 1)	$F_{(q-1),(p-1)(q-1)}$ = MSSB/MSSE
Error	(p - 1) (q - 1)	$SSE = \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$	$MSSE = \frac{SSE}{(p-1)(q-1)}$	
Total	pq - 1	$TSS = \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{..})^2$		

If both the null hypotheses are rejected then the means of the different levels of factors A and B can be tested by the method of multiple comparisons given by Tukkey.

7.5.1 General Computation Procedure

For analysing the two-way classified data, with one observation per cell, one has to follow the following procedure:

1. Calculate G = Grand Total = Total of all observations = $\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij}$
2. Find the N = The number of observations
3. Find Correction Factor(CF) = G^2/N
4. Calculate Raw Sum Squares(RSS) = $\sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij}^{\ 2}$
5. Compute Total Sum of Squares (TSS) = RSS - CF
6. Calculate Sum of Squares due to Factor A (SSA)

   $= y_{1.}^{\ 2}/q + y_{2.}^{\ 2}/q + ... + y_{i.}^{\ 2}/q + ... + y_{p.}^{\ 2}/q - CF$
7. Compute Sum of Squares due to Factor B (SSB)

   $= y_{.1}^{\ 2}/p + y_{.2}^{\ 2}/p + ... + y_{.j}^{\ 2}/p + ... + y_{.q}^{\ 2}/p - CF$
8. Compute Sum of Squares due to Error (SSE) =TSS – SSA - SSB
9. Compute MSSA = SSA/df,

   MSSB = SSB/df

   MSSE = SSE/df
10. Find F_A = MSSA/MSSE

    and F_B = MSSB/MSSE
11. Compare the calculated value of F_A to tabulated value of F_A; if calculated value is greater than the tabulated value then reject the hypothesis H_{0A}, otherwise it may be accepted.
12. Compare the calculated value of F_B to tabulated value of F_B; if calculated value is greater than the tabulated value then reject the hypothesis H_{0B}, otherwise it may be accepted.

7.6 DEGREES OF FREEDOM OF VARIOUS SUM OF SQUARES

Total sum of squares (TSS) is calculated from pq observations of the form $(y_{ij} - \bar{y}_{..})$ which carry $(pq-1)$degrees of freedom (df), one degree of freedom being lost because of the linear constraints $\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{..}) = 0$.

Similarly, the sum of squares due to the levels of factor A $\sum_{i=1}^{p} q(\bar{y}_{i.} - \bar{y}_{..})^2$ =SSA will have (p-1) degrees of freedom, since $\sum_{i=1}^{p} q(\bar{y}_{i.} - \bar{y}_{..}) = 0$.

The sum of squares due to the levels of factor B (SSB) = $\sum_{i=1}^{q} p(\bar{y}_{.j} - \bar{y}_{..})^2$ will have (q - 1) degrees of freedom, since $\sum_{i=1}^{p} p(\bar{y}_{.j} - \bar{y}_{..}) = 0$.

The error sum of squares (SSE) = $\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$ will have (p - 1) (q - 1) degrees of freedom (df). Since it is based on pq quantities which are subject to (p + q - 1) constraints or which can also be obtained by subtracting the degrees of freedom of sum of squares due to factor A and degrees of freedom of sum of squares due to factor B from the degrees of freedom for total sum of squares.

Hence, df of TSS = df (SSA) + df (SSB) + df (SSE)

$(pq-1) = (p-1) + (q-1) + (p-1)(q-1)$

7.7 EXPECTATIONS OF VARIOUS SUM OF SQUARES

For obtaining the valid test statistic, the expected value of various sum of squares should be obtained:

7.7.1 Expectation of Sum of Squares due to Factor A

We have $SSA = q\sum_{i=1}^{p}(\bar{y}_{i.} - \bar{y}_{..})^2$

$$E(SSA) = E\left[q\sum_{i=1}^{p}(\bar{y}_{i.} - \bar{y}_{..})^2\right]$$

$$E(SSA) = q\,E\left[\sum_{i=1}^{p}(\bar{y}_{i.} - \bar{y}_{..})^2\right] \quad \ldots (2)$$

We know that $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$

So $\bar{y}_{i.} = \frac{1}{q}\sum_{j=1}^{q} y_{ij} = \frac{1}{q}\left[\sum_{j=1}^{q}(\mu + \alpha_i + \beta_j + e_{ij})\right]$

or $\bar{y}_{i.} = \frac{1}{q}\left[q\mu + q\alpha_i + \sum_{j=1}^{q}\beta_j + \sum_{j=1}^{q} e_{ij}\right]$

$$= \mu + \alpha_i + \bar{e}_{i.}$$

Similarly, the value of $\bar{y}_{..}$ will be

$$\bar{y}_{..} = \mu + \bar{e}_{..}$$

Substituting the values of $\bar{y}_{i.}$ and $\bar{y}_{..}$ in equation (2), we have

$$E(SSA) = q\,E\left[\sum_{i=1}^{p}(\mu + \alpha_i + \bar{e}_{i.} - \mu - \bar{e}_{..})^2\right]$$

$$= q\,E\left[\sum_{i=1}^{p}(\alpha_i + \bar{e}_{i.} - \bar{e}_{..})^2\right]$$

$$= q\,E\left[\sum_{i=1}^{p}\alpha_i^2 + \sum_{i=1}^{p}(\bar{e}_{i.} - \bar{e}_{..})^2 + 2\sum_{i=1}^{p}\alpha_i(\bar{e}_{i.} - \bar{e}_{..})\right]$$

$$= q\left[\sum_{i=1}^{p}\alpha_i^2 + \sum_{i=1}^{p}E(\bar{e}_{i.} - \bar{e}_{..})^2 + 2\sum_{i=1}^{p}\alpha_i E(\bar{e}_{i.} - \bar{e}_{..})\right]$$

Since $e_{ij} \sim \text{iidN}(0, \sigma_e^2)$, so $E(e_{ij}) = 0 \quad \text{Var}(e_{ij}) = \sigma_e^2$

$$\bar{e}_{i.} \sim \text{iidN}\left(0, \frac{\sigma_e^2}{q}\right);\ E(\bar{e}_{i.}) = 0 \quad \text{Var}(\bar{e}_{i.}) = \frac{\sigma_e^2}{q}$$

and $$\bar{e}_{..} \sim \text{iidN}\left(0, \frac{\sigma_e^2}{pq}\right);\ E(\bar{e}_{..}) = 0 \quad \text{Var}(\bar{e}_{..}) = \frac{\sigma_e^2}{pq}$$

So $$E(SSA) = q\sum_{i=1}^{p}\alpha_i^2 + q\sum_{i=1}^{p}E(\bar{e}_{i.} - \bar{e}_{..})^2 + 0$$

Because $$E(\bar{e}_{i.}) = E(\bar{e}_{..}) = 0$$

or $$E(SSA) = q\sum_{i=1}^{p}\alpha_i^2 + qE\sum_{i=1}^{p}(\bar{e}_{i.}^2 + \bar{e}_{..}^2 - 2\bar{e}_{i.}\bar{e}_{..})$$

$$= q\sum_{i=1}^{p}\alpha_i^2 + qE\left[\sum_{i=1}^{p}\bar{e}_{i.}^2 + p\,\bar{e}_{..}^2 - 2\,\bar{e}_{..}\sum_{i=1}^{p}\bar{e}_{i.}\right]$$

$$= q\sum_{i=1}^{p}\alpha_i^2 + qE\left[\sum_{i=1}^{p}\bar{e}_{i.}^2 - p\,\bar{e}_{..}^2\right] \quad \text{because } \bar{e}_{..} = \frac{1}{p}\sum_{i=1}^{p}\bar{e}_{i.}$$

$$= q\sum_{i=1}^{p}\alpha_i^2 + q\left\{\sum_{i=1}^{p}E(\bar{e}_{i.}^2) - pE(\bar{e}_{..}^2)\right\}$$

$$= q\sum_{i=1}^{p}\alpha_i^2 + q\left(p\frac{\sigma_e^2}{q} - p\,\frac{\sigma_e^2}{pq}\right)$$

$$E(SSA) = q\sum_{i=1}^{p}\alpha_i^2 + (p-1)\sigma_e^2$$

$$E\left[\frac{SSA}{p-1}\right] = \frac{q}{(p-1)}\ \sum_{i=1}^{p}\alpha_i^2 + \sigma_e^2$$

Under H_{0A}, MSSA provides an unbiased estimate of σ_e^2.

7.7.2 Expectation of Sum of Squares due to Factor B

Similarly, the E (SSB) can be obtained

$$SSB = p\sum_{j=1}^{q}(\bar{y}_{.j} - \bar{y}_{..})^2$$

$$E(SSB) = E\left[p\sum_{j=1}^{q}(\bar{y}_{.j} - \bar{y}_{..})^2\right]$$

$$= p\,E\left[\sum_{j=1}^{q}\left(\bar{y}_{.j}-\bar{y}_{..}\right)^2\right] \qquad \ldots (3)$$

We know that $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$

So $$\bar{y}_{.j} = \frac{1}{p}\sum_{i=1}^{p} y_{ij} = \frac{1}{p}\left[\sum_{i=1}^{p}\left(\mu + \alpha_i + \beta_j + e_{ij}\right)\right]$$

or $$\bar{y}_{.j} = \frac{1}{p}\left[p\mu + \sum_{i=1}^{p}\alpha_i + p\beta_j + \sum_{i=1}^{p} e_{ij}\right]$$

$$= \mu + \beta_j + \bar{e}_{.j}$$

Similarly, the value of $\bar{y}_{..}$ will be

$$\bar{y}_{..} = \mu + \bar{e}_{..}$$

Substituting the values of $\bar{y}_{.j}$ and $\bar{y}_{..}$ in equation (3), we have

$$E\,(SSB) = p\,E\left[\sum_{j=1}^{q}\left(\mu + \beta_j + \bar{e}_{.j} - \mu - \bar{e}_{..}\right)^2\right]$$

$$= p\,E\left[\sum_{j=1}^{q}\left(\beta_j + \bar{e}_{.j} - \bar{e}_{..}\right)^2\right]$$

$$= p\,E\left[\sum_{j=1}^{q}\beta_j^2 + \sum_{j=1}^{q}\left(\bar{e}_{.j} - \bar{e}_{..}\right)^2 + 2\sum_{j=1}^{q}\beta_j\left(\bar{e}_{.j} - \bar{e}_{..}\right)\right]$$

$$= p\left[\sum_{j=1}^{q}\beta_j^2 + E\sum_{j=1}^{q}\left(\bar{e}_{.j} - \bar{e}_{..}\right)^2 + 2E\sum_{j=1}^{q}\beta_j\left(\bar{e}_{.j} - \bar{e}_{..}\right)\right]$$

Since, $e_{ij} \sim$ iid $N(0, \sigma_e^2)$, so $E\left(e_{ij}\right) = 0$ $\quad Var\left(e_{ij}\right) = \sigma_e^2$

$$\bar{e}_{.j} \sim \text{iid}\ N\left(0, \frac{\sigma_e^2}{p}\right) \quad E\left(\bar{e}_{.j}\right) = 0 \quad Var\left(\bar{e}_{.j}\right) = \frac{\sigma_e^2}{p}$$

and $$\bar{e}_{..} \sim \text{iid}\ N\left(0, \frac{\sigma_e^2}{pq}\right) \quad E\left(\bar{e}_{..}\right) = 0 \quad Var\left(\bar{e}_{..}\right) = \frac{\sigma_e^2}{pq}$$

Because $E\left(\bar{e}_{.j}\right) = E\left(\bar{e}_{..}\right) = 0$

Therefore, $$E(SSB) = p\sum_{j=1}^{q}\beta_j^2 + pE\sum_{j=1}^{q}\left(\bar{e}_{.j}^2 + \bar{e}_{..}^2 - 2\bar{e}_{.j}\bar{e}_{..}\right)$$

$$= p\sum_{j=1}^{q}\beta_j^2 + pE\left[\sum_{j=1}^{q}\bar{e}_{.j}^2 + q\bar{e}_{..}^2 - 2\bar{e}_{..}\sum_{j=1}^{q}\bar{e}_{.j}\right]$$

$$= p\sum_{j=1}^{q}\beta_j^2 + pE\left[\sum_{j=1}^{q}\bar{e}_{.j}^2 - q\bar{e}_{..}^2\right] \quad \text{because } \bar{e}_{..} = \frac{1}{q}\sum_{j=1}^{q}\bar{e}_{.j}$$

$$= p\sum_{j=1}^{q}\beta_j^2 + p\left[E\left(\sum_{j=1}^{q}\bar{e}_{.j}^2\right) - qE\left(\bar{e}_{..}^2\right)\right]$$

$$= p\sum_{j=1}^{q}\beta_j^2 + p\left(q\frac{\sigma_e^2}{p} - q\,\frac{\sigma_e^2}{pq}\right)$$

$$E(SSB) = p\sum_{j=1}^{q}\beta_j^2 + (q-1)\sigma_e^2$$

$$E\left[\frac{SSB}{q-1}\right] = \frac{p}{(q-1)}\sum_{j=1}^{q}\beta_j^2 + \sigma_e^2$$

Under H_{0B}, the MSSB is unbiased estimate of σ_e^2

7.7.3 Expectation of Sum of Squares Due to Error

$$E(SSE) = E\left(\sum_{i=1}^{p}\sum_{j=1}^{q}\left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2\right)$$

$$= E\left(\sum_{i=1}^{p}\sum_{j=1}^{q}\left(\mu + \alpha_i + \beta_j + e_{ij} - \mu - \alpha_i - \bar{e}_{i.} - \mu - \beta_j - \bar{e}_{.j} + \mu + \bar{e}_{..}\right)^2\right)$$

$$= E\left(\sum_{i=1}^{p}\sum_{j=1}^{q}\left(e_{ij} - \bar{e}_{i.} - \bar{e}_{.j} + \bar{e}_{..}\right)^2\right)$$

$$= E\Bigg[\sum_{i=1}^{p}\sum_{j=1}^{q}e_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\bar{e}_{i.}^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\bar{e}_{.j}^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\bar{e}_{..}^2$$
$$- 2\sum_{i=1}^{p}\sum_{j=1}^{q}e_{ij}\bar{e}_{i.} - 2\sum_{i=1}^{p}\sum_{j=1}^{q}e_{ij}\bar{e}_{.j} + 2\sum_{i=1}^{p}\sum_{j=1}^{q}e_{ij}\bar{e}_{..}$$
$$+ 2\sum_{i=1}^{p}\sum_{j=1}^{q}\bar{e}_{i.}\bar{e}_{.j} - 2\sum_{i=1}^{p}\sum_{j=1}^{q}\bar{e}_{i.}\bar{e}_{..} - 2\sum_{i=1}^{p}\sum_{j=1}^{q}\bar{e}_{.j}\bar{e}_{..}\Bigg]$$

$$= E\Bigg[\sum_{i=1}^{p}\sum_{j=1}^{q}e_{ij}^2 + q\sum_{i=1}^{p}\bar{e}_{i.}^2 + p\sum_{j=1}^{q}\bar{e}_{.j}^2 + pq\bar{e}_{..}^2 - 2q\sum_{i=1}^{p}\bar{e}_{i.}^2$$
$$- 2p\sum_{j=1}^{q}\bar{e}_{.j}^2 + 2pq\bar{e}_{..}^2 + 2pq\bar{e}_{..}^2 - 2pq\bar{e}_{..}^2 - 2pq\bar{e}_{..}^2\Bigg]$$

$$= E\left[\sum_{i=1}^{p}\sum_{j=1}^{q}e_{ij}^2 - q\sum_{i=1}^{p}\bar{e}_{i.}^2 - p\sum_{j=1}^{q}\bar{e}_{.j}^2 + pq\bar{e}_{..}^2\right]$$

$$E(SSE) = \sum_{i=1}^{p}\sum_{j=1}^{q}E(e_{ij}^2) - q\sum_{i=1}^{p}E(\bar{e}_{i.}^2) - p\sum_{j=1}^{q}E(\bar{e}_{.j}^2) + pqE(\bar{e}_{..}^2)$$

Since $e_{ij} \sim$ iid $N(0, \sigma_e^2)$, so $E(e_{ij}) = 0$ $Var(e_{ij}) = \sigma_e^2$

so $E(SSE) = pq\,\sigma_e^2 - qp.\frac{\sigma_e^2}{q} - pq\frac{\sigma_e^2}{p} + pq.\frac{\sigma_e^2}{pq}$

$= pq\,\sigma_e^2 - p\,\sigma_e^2 - q\,\sigma_e^2 + \sigma_e^2$

$= (pq - p - q + 1)\,\sigma_e^2$

$E(SSE) = (p-1)(q-1)\,\sigma_e^2$

$$E\left(\frac{SSE}{(p-1)(q-1)}\right) = \sigma_e^2$$

or $E(MSSE) = \sigma_e^2$

Hence, mean sum of squares due to error is unbiased estimate of σ_e^2.

Example 1: Future group wishes to enter the frozen shrimp market. They contract a researcher to investigate various methods of groups of shrimp in large tanks. The researcher suspects that temperature and salinity are important factor influencing shrimp yield and conducts a two-way analysis of variance with their levels of temperature and salinity. That is each combination of yield for each (for identical gallon tanks) is measured.The recorded yields are given in the following chart:

Categorical variable Salinity (in pp)

Temperature	700	1400	2100	Total	Mean
60^0 F	3	5	4	12	4
70^0 F	11	10	12	33	11
80^0 F	16	21	17	54	18
Total	30	36	33	99	11

Compute the ANOVA table for the model.

Solution: Since in each call there is one observation. So we will use the model

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

where, y_{ij} is the yield corresponding to i^{th} temperature and j^{th} salinity, μ is the general mean, α_i is the effect due to i^{th} temperature, β_j is the effect due to j^{th} salinity and $e_{ij} \sim$ i i d N $(0, \sigma^2)$

Test of hypothesis that there is a significant difference in shrimp yield due to difference in the temprature or

$H_{0A}: \alpha_1 = \alpha_2 = \alpha_3$ or $H_{0A}: \alpha_i = 0$ for all i=1, 2, 3

$H_{1A}: \alpha_1 \neq \alpha_2 \neq \alpha_3 \neq 0$ or $H_{1A}: \alpha_i \neq 0$ at least one i

Similarly, there is a significant difference in shrimp yield due to difference in the salinity or

$H_{0B}: \beta_1 = \beta_2 = \beta_3$ or $H_{0B}: \beta_j = 0$ for all j=1, 2, 3

$H_{1B}: \beta_1 \neq \beta_2 \neq \beta_3 \neq 0$ or $H_{1B}: \beta_j \neq 0$ at least one j

The computations are as follows:

Grand Total = G = 99

No. of observations = N = 9

Correction Factor (CF) = $(99 \times 99) / 9 = 1089$

Raw Sum of Squares (RSS) = 1401

Total Sum of Squares (TSS) = RSS – CF = 1401 - 1089 = 312

Sum of Squares due to Temperature (SST)

$= (12)^2/3 + (33)^2/3 + (54)^2/3 - 1089$

$=(144 + 1089 + 2916)/3 - 1089$

$=(4149/3) - 1089$

$=1383 - 1089 = 294$

Sum of Squares due to Salinity (SSS)

$= (30)^2/3 + (36)^2/3 + (33)^2/3 - 1089$

$= (900 + 1296 + 1089)/3 - 1089$

$= (3285/3) - 1089$

$=1095 - 1089 = 6$

Sum of Squares due to Error = TSS – SST - SSS

$= 312 - 294 - 6 = 12$

ANOVA Table for Two-Way Classified Data

Sources of Variation	DF	SS	MSS	F-Test
Due to Temperature	2	294	147	F_T = MSST/MSSE =147/3 = 49
Due to Salinity	2	6	3	F_S = 3/3 = 1
Due to error	4	12	3	
Total	8	312		

Since tabulated value of $F_{2,4}$ at 5% level of significance is 10.649 which is less than the calculated $F_{T(2,4)} = 49$, for testing the significant difference in shrimp yield due to differences in levels of temperature. So, H_{0A} is rejected. Hence, there are differences in shrimp yield due to temperature at 5% level of significance.

Since tabulated value of $F_{2,4}$ at 5% level of significance is 10.649 which is greater than the calculated $F_{S(2,4)}=1$, for testing the significant difference in shrimp yield due to difference in the level of salinity. So, H_{0B} is accepted. Hence there is no any difference in shrimp yield due the salinity level at 5% level of significance.

In summary we conclude that $\beta_j = 0$ for j =1, 2, 3 and that $\alpha_i \neq 0$ for at each one value of i. So this model implies that temperature is a more important factor than salanity in influencing shrimp yield. Difference in the level of salanity appears to have no effect on shrimp at all.

H_{0A}: $\alpha_i \neq 0$ i= 1, 2, 3, then we have to test pairwise. So, multiple comparison tests will be applied for different levels of temperature so the null hypothesis will be

$H_{0A} : \alpha_i = \alpha_j \quad i \neq j = 1, 2, 3$

$H_{1A} : \alpha_i \neq \alpha_j$

So, if $(\bar{y}_{1.} - \bar{y}_{2.}) > t_{(p-1)(q-1)}\sqrt{\frac{2MSSE}{3}}$

Then, H_{0A} is rejected

So $(\bar{y}_{1.} - \bar{y}_{2.}) = 7, (\bar{y}_{1.} - \bar{y}_{3.}) = 14, (\bar{y}_{2.} - \bar{y}_{3.}) = 7$

and the value $t_{(p-1)(q-1)} \sqrt{\frac{2MSSE}{3}} = 2.1776 \times \sqrt{\frac{2 \times 3}{3}} = 3.93$

Since, all $(\bar{y}_{i.} - \bar{y}_{j.}) > 3.93$. So, we conclude that there is a significance difference among the yield of shrimp due to different levels of temperature.

E 1) An experiment was conducted to determine the effect of different data of planting and different methods of planting on the field of sugar-cane. The data below show the fields of sugar-cane for four different data and the methods of planting:

Data of Planting

Method of Planting	October	November	February	March
I	7.10	3.69	4.70	1.90
II	10.29	4.79	4.50	2.64
III	8.30	3.58	4.90	1.80

Carry out an analysis of the above data.

E 2) A researcher wants to test four diets A, B, C, D on growth rate in mice. These animals are divided into 3 groups according to their weights. Heaviest 4, next 4 and lightest 4 are put in Block I, Block II and Block III respectively. Within each block, one of the diets is given at random to the animals. After 15 days, increase in weight is noted, which given in the following table:

Blocks	Treatments/Diets			
	A	B	C	D
I	12	8	6	5
II	15	12	9	6
III	14	10	8	5

Perform a two-way ANOVA to test whether the data indicate any significant difference between the four diets due to different blocks.

7.8 SUMMARY

In this unit, we have discussed:

1. The two-way analysis of variance model;
2. The basic assumptions in two-way analysis of variance;
3. Test of hypothesis using two-way analysis of variance; and
4. Expectations of various sum of squares.

7.9 SOLUTIONS /ANSWERS

E1) H_{0A} : There is no difference among the different method of planting.

$$H_{0A}: \alpha_i = \alpha_2 = \alpha_3$$

Against, $H_{1A}: \alpha_i \neq \alpha_2 \neq \alpha_3$

H_{0B} : There is no any difference among the different data of planting.

$$H_{0B}: \beta_i = \beta_2 = \beta_3 = \beta_4$$

$$H_{1B}: \beta_i \neq \beta_2 \neq \beta_3 \neq \beta_4$$

$G = \sum\sum y_{ij}$ = Grand Total = Total of all observations

$$= 7.10 + 3.69 + 4.70 + 1.90 + 10.29 + 4.79 + 4.58 + 2.64 + 8.30 + 3.58 + 4.90 + 1.80$$

$$= 58.28$$

N = No. of observations = 12

$$\text{Correction Factor (CF)} = \frac{G^2}{N} = \frac{58.28 \times 58.28}{12} = 283.0465$$

Raw Sum of Squares (RSS)

$$= (7.10)^2 + (3.69)^2 + (4.70)^2 + (1.90)^2 + (10.29)^2 + (4.79)^2 + (4.58)^2 + (2.64)^2 + (8.30)^2 + (3.58)^2 + (4.90)^2 + (1.80)^2$$

$$= 355.5096$$

Total Sum of Squares (TSS) = RSS – CF

$$= 355.5096 - 283.0465 = 72.4631$$

Sum of Squares due to Data of Planting (SSD)

$$= \frac{D_1^2}{3} + \frac{D_2^2}{3} + \frac{D_3^2}{3} + \frac{D_4^2}{3} - CF$$

$$= \frac{(25.69)^2}{3} + \frac{(12.06)^2}{3} + \frac{(14.18)^2}{3} + \frac{(6.35)^2}{3} - 283.0465$$

$$= 65.8917$$

Sum of Squares due to Method of Planting (SSM)

$$= \frac{M_1^2}{4} + \frac{M_2^2}{4} + \frac{M_3^2}{4} - CF$$

$$= \frac{(17.39)^2}{4} + \frac{(22.31)^2}{4} + \frac{(15.58)^2}{4} - 283.0465$$

$$= 286.3412 - 283.0465 = 3.2947$$

Sum of Squares due to Error (SSE) = TSS - SSD - SSM

$$= 72.4631 - 3.2947 - 65.8917$$

$$= 3.2767$$

ANOVA Table

Source of Variation	SS	DF	MSS	F
Between Method	3.2947	2	$\frac{3.2947}{2}=1.6473$	$F_1=\frac{1.6473}{0.5461}=3.02$
Between Data	65.8917	3	$\frac{65.8917}{3}=21.9639$	$F_2=\frac{21.9639}{0.5461}=40.22$
Due to Errors	3.2767	6	$\frac{3.2767}{6}=0.5461$	
Total	72.4631	11		

The tabulated value of $F_{2,6}$ at 5% level of significance is 5.14 which is greater than the calculated value of F_M (3.02) so H_{0A} is accepted. So, we conclude that there is no significant difference among the different methods of planting.

The tabulated value of $F_{3,6}$ at 5% level of significance is 4.76 which is less than calculated value of F_D (40.22). So we reject the null hypothesis H_{0B}. Hence there is a significant difference among the data of planting. In all, we conclude that the different methods of planting affect the mean field of sugar-cane in the same manner. But the mean field differs with different data of planting.

If the four data of planting, included in the above experiment, be the only data in which we are interested. Then the next question that arises is: which one of the four data will give us the maximum mean field? To answer this question, we complete the critical difference (CD) at, say, the 5% level of significance,

$$CD = t_{\text{error df}}\sqrt{\frac{2MSSE}{3}} = 2.447\times\sqrt{0.3641} = 1.48$$

The mean field differences of the four data of planting are given as:

$$\left|\overline{D}_1-\overline{D}_2\right| = |8.56-4.02| = 4.54$$

$$\left|\overline{D}_1-\overline{D}_3\right| = |8.56-4.73| = 3.77$$

$$\left|\overline{D}_1-\overline{D}_4\right| = |8.56-2.12| = 6.44$$

$$\left|\overline{D}_2-\overline{D}_3\right| = |4.02-4.73| = 0.71$$

$$\left|\overline{D}_2-\overline{D}_4\right| = |4.02-2.12| = 1.90$$

$$\left|\overline{D}_3-\overline{D}_4\right| = |4.73-2.12| = 2.61$$

From above we conclude that there is no any significant difference between the data of planting November and February. But there is a significant difference between October, November, December and February. But the mean of October is the maximum. So, we can release the data of planting in the month of October.

E2) Null hypotheses are

H_{01}: There is no significant difference between mean effects of diets.

H_{02}: There is no significant difference between mean effects of different blocks.

Against the alternative hypothesis

H_{11}: There is significant difference between mean effects of diets

H_{12}: There is significant difference between mean effects of different blocks.

Blocks	**Treatments/Diets**				
	A	B	C	D	**Totals**
I	12	8	6	5	31 $T_{.1}$
II	15	12	9	6	42 $T_{.2}$
III	14	10	8	5	37 $T_{.3}$
Totals	41 $T_{1.}$	30 $T_{2.}$	23 $T_{3.}$	16 $T_{4.}$	110 Grand Total

Squares of observations

Blocks	**Treatments/Diets**				**Totals**
	A	B	C	D	
I	144	64	36	25	269
II	225	144	81	36	486
III	196	100	64	25	385
Totals	565	308	181	86	1140

Grand Total = $G = \sum\sum y_{ij} = 110$

Correction Factor (CF) = $\frac{G^2}{n} = \frac{(110)^2}{12} = 1008.3333$

Raw Sum of Squares (RSS) = $\sum\sum y_{ij}^{\ 2} = 1140$

Total Sum Squares (TSS) = RSS - CF = 1140 - 1008.3333

= 131.6667

Sum of Squares due to Treatments/ Diets (SST)

$$= \frac{T_{1.}^2}{3} + \frac{T_{2.}^2}{3} + \frac{T_{3.}^2}{3} + \frac{T_{4.}^2}{3} - CF$$

$$= \frac{1}{3}\left[(41)^2 + (30)^2 + (23)^2 + (16)^2\right] - 1008.3333$$

$$= \frac{1}{3}\left[1681 + 900 + 529 + 256\right] - 1008.3333$$

$$= 1122 - 1008.3333$$

$$= 113.6667$$

Sum of Squares due to Block (SSB) = $\frac{1}{4}\left(T_{.1}^2 + T_{.2}^2 + T_{.3}^2\right) - CF$

$$= \frac{1}{4}\left[259 + 486 + 385\right] - 1008.3333$$

$$= 1023.5 - 1008.3333 = 15.1667$$

Sum of Squares due to Errors (SSE) = TSS – SST – SSB

$$= 131.6667 - 113.6667 - 15.1667$$

$$= 2.8333$$

Mean Sum of Squares due to Treatments (MSST)

$$= \frac{SST}{df} = \frac{113.6667}{3} = 37.8889$$

Mean Sum of Squares due to Blocks (MSSB)

$$= \frac{SSB}{df} = \frac{15.1667}{2} = 7.58335$$

Mean Sum of Squares due to Errors (MSSE)

$$= \frac{SSE}{df} = \frac{2.8333}{6} = 0.4722$$

ANOVA Table

Source of Variation	SS	DF	MSS	F
Between Treatments/ Diets	113.6667	3	$\frac{113.6667}{3} = 37.8889$	$F_1 = \frac{37.8889}{0.4722} = 80.2391$
Between Blocks	15.1667	2	$\frac{15.1667}{2} = 7.58335$	$F_2 = \frac{7.58335}{0.4722} = 160.596$
Due to Errors	2.8333	6	$\frac{2.8333}{6} = 0.4722$	
Total	131.6667	11		

Tabulated F at 5% level of significance with (3, 6) degrees of freedom is 4.76 & tabulated F at 5% level of significance with (2, 6) degrees of freedom is 5.14

Conclusion: Since calculated F_1 > tabulated F, so we reject H_{01} and conclude that there is significant difference between mean effect of diets.

Also calculated F_2 is greater than tabulated F, so we reject H_{02} and conclude that there is significant difference between mean effect of different blocks.

UNIT 8 TWO-WAY ANOVA WITH m OBSERVATIONS PER CELL

Structure

8.1 Introduction
Objectives
8.2 ANOVA Model for Two-way Classified Data with m Observations per Cell
8.3 Basic Assumptions
8.4 Estimation of Parameters
8.5 Test of Hypothesis
8.6 Degrees of Freedom of Various Sum of Squares
8.7 Expectations of Various Sum of Squares
8.8 Summary
8.9 Solutions/Answers

8.1 INTRODUCTION

In the analysis of variance technique, if explanatory variable is only one and different levels of independent variable is under consideration then it is called one-way analysis of variance and a test of hypothesis is developed for the equality of several mean of different levels of a factor/independent variable/ explanatory variable. But if we are interested to consider two independent variables for analysis in place of one, and able to perform the two hypotheses for the levels of these factors independently (there is no interaction between these two factors). The above analysis has been given in the Units 6 and 7 respectively. But if we are interested to test the interaction between two factors and we have repeated observations then the two-way analysis of variance with m observation per cell is considered. If there are exactly same numbers of observations in the cell then it is called balance.

In this unit, a mathematical model for two-way classified data with m-observations per cell is given in Section 8.2. The basic assumptions are given in Section 8.3 whereas the estimation of parameters is given in Section 8.4. Test of hypothesis for two-way ANOVA is explained in Section 8.5 and degrees of freedom of various sum of squares are described in Section 8.6. The expected values of sum of squares for two factors and their interactions are derived in Section 8.7.

Objectives

After studying this unit, you would be able to

- describe the ANOVA model for two-way classified data with m observations per cell;
- describe the basic assumptions for the given model;
- obtain the estimates of the parameters of the given model;

- describe the test of hypothesis for two-way classified data with m observations per cell;
- derive the expectations of the various sum of squares; and
- perform to test the hypothesis for two-way classified data with m observations per cell.

8.2 ANOVA MODEL FOR TWO-WAY CLASSIFIED DATA WITH m OBSERVATIONS PER CELL

In Unit 7, it was seen that we cannot obtain an estimate of, or make a test for the interaction effect in the case of two-way classified data with one observation per cell. This is possible, however, if some or all of the cells contain more than one observations. We shall assume that there is an equal number of (m) observations in each cell. The m observations in the $(i, j)^{th}$ cell will be denoted $y_{ij1}, y_{ij2}, \ldots, y_{ijm}$. Thus, y_{ijk} is the k^{th} observation for i^{th} level of factor A and j^{th} level of factor B, i = 1, 2, …, p; j = 1, 2, …,q & k = 1, 2, …, m.

The mathematical model

$$y_{ijk} = \mu_{ij} + e_{ijk}$$

where μ_{ij} is the true value for the $(i, j)^{th}$ cell and e_{ijk} is the error. e_{ijk} are assumed to be independently identical normally distributed, each with mean zero and variance σ_e^2. The table of observations can be displayed as follows:

A/B	B_1	B_2	B_j	…	B_q	Total	Total
A_1	y_{111}	y_{121}	y_{1j1}	…	y_{1q1}	$y_{1.1}$	$y_{1..}$
	y_{112}	y_{122}	y_{1j2}	…	y_{1q2}	$y_{1.2}$	
	.	.	.		.	.	
	.	.	.		.	.	
	.	.	.		.	.	
	y_{11m}	y_{12m}	y_{1jm}	…	y_{1qm}	$y_{1.m}$	
A_2	y_{211}	y_{221}	y_{2j1}	…	y_{2q1}	$y_{2.1}$	$y_{2..}$
	y_{212}	y_{222}	y_{2j2}	…	y_{2q2}	$y_{2.2}$	
	.	.	.		.	.	
	.	.	.		.	.	
	.	.	.		.	.	
	y_{21m}	y_{22m}	y_{2jm}	…	Y_{2qm}	$y_{2.m}$	
.	y_{i11}	y_{i21}	y_{ij1}	…	y_{iq1}	$y_{i.1}$	$y_{i..}$
.	y_{i12}	y_{i22}	y_{ij2}	…	y_{iq2}	$y_{i.2}$	
.	.	.	.	…	.	.	
A_i	.	.	.	…	.	.	
.	.	.	.	…	.	.	
.	y_{i1m}	y_{i2m}	y_{ijm}	…	y_{iqm}	$y_{i.m}$	
A_p	y_{p11}	y_{p21}	y_{pj1}	…	y_{pq1}	$y_{p.1}$	$y_{p..}$
	y_{p12}	y_{p22}	y_{pj2}	…	y_{pq2}	$y_{p.2}$	
	.	.	.		.	.	
	.	.	.		.	.	
	.	.	.		.	.	
	y_{p1m}	y_{p2m}	y_{pjm}	…	y_{pqm}	$y_{p.m}$	
Total	$y_{.1.}$	$y_{.2.}$	$y_{.j.}$	…	$y_{.q.}$	$y_{...}$	

The model can be written as

$$y_{ijk} = \mu + (\mu_{i.} - \mu) + (\mu_{.j} - \mu) + (\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu) + e_{ijk}$$

$$= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

where, μ is general mean effect, α_i is the effect of i^{th} level of the factor A, β_j is the effect of j^{th} level of factor B, $(\alpha\beta)_{ij}$ is the interaction effect between i^{th} level of A factor and j^{th} level of B factor.

$$\sum_{i=1}^{p} \alpha_i = 0, \sum_{j=1}^{q} \beta_j = 0, \sum_{i=1}^{p} (\alpha\beta)_{ij} = 0, \sum_{j=1}^{q} (\alpha\beta)_{ij} = 0$$

where,

$y_{...}$ = Sum of all the observations.

$y_{i..}$ = Total of all observations in the i^{th} level of factor A

$y_{.j.}$ = Total of all observations in the j^{th} level of factor B.

8.3 BASIC ASSUMPTIONS

Following assumptions should be followed for valid and reliable test procedure for testing of hypothesis as well as for estimation of parameters

1. All the observations y_{ijk} are independent.
2. Different effects are additive in nature.
3. e_{ijk} are independent and identicaly distributed as normal with mean zero and constant variance σ_e^2.

8.4 ESTIMATION OF PARAMETERS

The least square estimates for various effects, obtained by minimizing the residual sum of squares

$$E = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{m} \left[y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij} \right]^2$$

by partially differentiating E with respect to μ, α_i (i=1, 2, ..., p), β_j (i=1, 2, ..., q) and $(\alpha\beta)_{ij}$ for all i = 1, 2, ..., p; j =1, 2, … , q and equating these equations equal to zero. These equations are called normal equations. Solution of these normal equations provide the estimates of these parameters [μ, α_i, β_j, $(\alpha\beta)_{ij}$].

$$\frac{\partial E}{\partial \mu} = -2 \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{m} \left[y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij} \right] = 0$$

$$\frac{\partial E}{\partial \alpha_i} = -2 \sum_{j=1}^{q} \sum_{k=1}^{m} \left[y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij} \right] = 0$$

$$\frac{\partial E}{\partial \beta_j} = -2\sum_{i=1}^{p}\sum_{k=1}^{m}\left[y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij}\right] = 0$$

$$\frac{\partial E}{\partial (\alpha\beta)_{ij}} = -2\sum_{k=1}^{m}\left[y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij}\right] = 0$$

These equations give, $\hat{\mu} = \dfrac{\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} y_{ijk}}{pqm} = \bar{y}_{...}$

$$\hat{\alpha}_i = \frac{\sum_{j=1}^{q}\sum_{k=1}^{m} y_{.jk}}{qm} - \hat{\mu} = \bar{y}_{i..} - \bar{y}_{...}$$

Similarly, $$\hat{\beta}_j = \frac{\sum_{i=1}^{p}\sum_{k=1}^{m} y_{i.k}}{pm} - \hat{\mu} = \bar{y}_{.j.} - \bar{y}_{...}$$

$$\hat{(\alpha\beta)}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$$

Substituting the values of $\hat{\mu}_i$, $\hat{\alpha}_i$, $\hat{\beta}_j$ and $(\hat{\alpha\beta})_{ij}$, in the model and then select the value of e_{ijk} such that both the sides are equal, so

$$y_{ijk} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})$$

or

$$y_{ijk} - \bar{y}_{...} = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})$$

Squaring and summing both the sides over i, j & k, then we get

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(y_{ijk} - \bar{y}_{...})^2 = mq\sum_{i=1}^{p}(\bar{y}_{i..} - \bar{y}_{...})^2 + mp\sum_{j=1}^{q}(\bar{y}_{.j.} - \bar{y}_{...})^2$$
$$+ m\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(y_{ijk} - \bar{y}_{ij.})^2$$

as usual product terms vanish.

Total Sum of Squares = Sum of Squares due to Factor A+ Sum of Squares due to Factor B + Sum of Squares due to Interaction A and B + Sum of Squares due to Error

or TSS = SSA + SSB + SSAB + SSE

8.5 TEST OF HYPOTHESIS

There are three hypotheses which are to be tested are as follows:

H_{0A}: $\alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$

H_{1A}: $\alpha_1 \neq \alpha_2 \neq \ldots \neq \alpha_p \neq 0$

H_{0B}: $\beta_1 = \beta_2 = \ldots = \beta_q = 0$

H_{1B}: $\beta_1 \neq \beta_2 \neq \ldots \neq \beta_q \neq 0$

H_{0AB}: $(\alpha\beta)_{ij} = 0$ for all i and j or A and B are independent to each other

H_{1AB}: $(\alpha\beta)_{ij} \neq 0$

The appropriate test statistics for testing the above hypothesis is:

$$F = \frac{SSA/(p-1)}{SSE/pq(m-1)} = \frac{MSSA}{MSSE}$$

If this value of F is greater than the tabulated value of F with [(p - 1), pq (m - 1)] df at α level of significance so we reject the null hypothesis, otherwise we may accept the null hypothesis.

Similarly, test statistics for second and third hypotheses are

$$F = \frac{SSB/(q-1)}{SSE/pq(m-1)} = \frac{MSSB}{MSSE}$$

$$F = \frac{SS\,AB/(p-1)(q-1)}{SSE/pq(m-1)} = \frac{MSSAB}{MSSE}$$

For practical point of view, first we should decide whether or not H_{0AB} can be rejected at an appropriate level of significance by using above F. If interaction effects are not significant i.e. the factor A and factor B are independent then we can find the best level of A and best level of B by multiple comparison method using t-test. On the other hand, if they are found to be significant, there may not be a single level of factor A and single level of factor B that will be the best in all situations. In this case, one will have to compare for each level of B at the different levels of A and for each level of A at the different levels of B.

The above analysis can be shown in the following ANOVA table:

ANOVA Table for Two-way Classified Data with m Observations per Cell

Sourses of Variation	**DF**	**SS**	**MSS**	**F**
Between the levels of A	p - 1	$SSA = mq\sum_{i=1}^{p}(\bar{y}_{i..} - \bar{y}_{...})^2$	MSSA = SSA / (p - 1)	F = MSSA / MSSE
Between the levels of B	q - 1	$SSB = mp\sum_{j=1}^{q}(\bar{y}_{.j.} - \bar{y}_{...})^2$	MSSB = SSB / (q - 1)	F = MSSB / MSSE
Interaction AB	(p - 1) (q - 1)	$SSAB = m\sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	MSSAB = SSAB / (p - 1)(q - 1)	F = MSS(AB) / MSSE
Error	pq (m - 1)	$TSS = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(y_{ijk} - \bar{y}_{ij.})^2$	MSSE = SSE / pq (m - 1)	
Total	mpq - 1			

Steps for Calculating Various Sums of Squares

1. Calculate G = Grand Total = Total of all observations = $\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} y_{ijk}$
2. Determine N = Numebr of observations.
3. Find Correction Factor (CF) = G^2/N
4. Raw Sum of Squares (RSS) = $\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} y_{ijk}^2$
5. Total Sum of Squares (TSS) = RSS - CF
6. Sum of Squares due to Factor A (SSA)
   $= \{y_{1..}^2/mq + y_{2..}^2/mq + ... + y_{i..}^2/mq + ... + y_{p..}^2/mq\} - CF$
7. Sum of Squares due to Factor B (SSB)
   $= \{y_{.1.}^2/mp + y_{.2.}^2/mp + \ldots + y_{.j.}^2/mp + ... + y_{.q.}^2/mp\} - CF$
8. Sum of Squares due to Means (SSM)
   $= \{y_{..1}^2/pq + y_{..2}^2/pq + \ldots + y_{..k}^2/pq + ... + y_{..m}^2/pq\} - CF$
9. Sum of Squares due to Interation AB(SSAB) = SSM – SSA - SSB
10. Sum of Squares due to Error (SSE) = TSS -SSA-SSB-SSAB
11. Calculate MSSA = SSA/df
12. Calculate MSSB = SSB/df
13. Calculate MSSAB = SS(AB)/df
14. Calculate MSSE = SSE/df
15. Calculate F_A = MSSA/MSSE ~ $F_{(p-1),\ pq(m-1)}$
16. Calculate F_B = MSSB/MSSE ~ $F_{(q-1),\ pq(m-1)}$
17. Calculate F_{AB} = MSS(AB)/MSSE ~ $F_{((p-1)(q-1),\ pq(m-1)}$

8.6 DEGREES OF FREEDOM OF VARIOUS SUM OF SQUARES

Total sum of squares (TSS) considers the pqm observations so the degrees of freedom for TSS are (pqm-1). One degree of freedom is lost due to the restriction that $\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}(y_{ijk} - \bar{y}_{...}) = 0.$

The degrees of freedom for sum of squares due to factor A is (p-1) because it has p levels. Similarly, the degrees of freedom for sum of squares due to factor B is (q-1) because it has q levels, under consideration. Sum of squares due to interaction of factors A and B is (p-1) (q-1) and the degrees of freedom for sum of squares due to errors is pq (m-1). Thus partitioning of degrees of freedom is as follows:

$$(mpq-1) = (p-1) + (q-1) + (p-1)(q-1) + pq(m-1)$$

which implies that the df are additive.

8.7 EXPECTATIONS OF VARIOUS SUM OF SQUARES

8.7.1 Expected Value of Sum of Squares due to Factor A

$$E\,(SSA) = E\left[mq\sum_{i=1}^{p}\left(\bar{y}_{i..} - \bar{y}_{...}\right)^2 \right]$$

Substituting the value of $\bar{y}_{i..}$ and $\bar{y}_{...}$ from the model, we get

$$= E\left[mq\sum_{i=1}^{p}\left(\alpha_i + \bar{e}_{i..} - \bar{e}_{...}\right)^2 \right]$$

or
$$E\,(SSA) = E\left[mq\sum_{i=1}^{p}\left\{\alpha_i^2 + \left(\bar{e}_{i..} - \bar{e}_{...}\right)^2 + 2\alpha_i\left(\bar{e}_{i..} - \bar{e}_{...}\right)\right\}\right]$$

$$= mq\sum_{i=1}^{p} E\left\{\alpha_i^2 + \left(\bar{e}_{i..} - \bar{e}_{...}\right)^2 + 2\alpha_i\left(\bar{e}_{i..} - \bar{e}_{...}\right)\right\}$$

$$= mq\sum_{i=1}^{p}\alpha_i^2 + mqE\left\{\sum_{i=1}^{p}\left(\bar{e}_{i..} - \bar{e}_{..}\right)^2\right\} + 0$$

$$\left[\text{Because } E\left(\bar{e}_{i..} - \bar{e}_{..}\right) = 0\right]$$

$$\therefore E(SSA) = mq\sum_{i=1}^{p}\alpha_i^2 + mq\,E\left[\sum_{i=1}^{p}\left(\bar{e}_{i..}\right)^2 - p\,\bar{e}_{...}^2\right]$$

$$= mq\sum_{i=1}^{p}\alpha_i^2 + mq\left[\sum_{i=1}^{p} E\left(\bar{e}_{i..}\right)^2 - p\,E\left(\bar{e}_{...}^2\right)\right]$$

$$= mq\sum_{i=1}^{p}\alpha_i^2 + mq\left[\sum_{i=1}^{p}\frac{\sigma_e^2}{mq} - p\,\frac{\sigma_e^2}{mpq}\right]$$

$$= mq\sum_{i=1}^{p}\alpha_i^2 + p\,\sigma_e^2 - \sigma_e^2$$

$$\therefore\ E\left(SSA\right) = mq\sum_{i=1}^{p}\alpha_i^2 + (p-1)\,\sigma_e^2$$

$$\Rightarrow E\left(\frac{SSA}{p-1}\right) = \frac{mq}{(p-1)}\sum_{i=1}^{p}\alpha_i^2 + \sigma_e^2$$

or
$$E\left(MSSA\right) = \sigma_e^2 + \frac{mq}{(p-1)}\sum_{i=1}^{p}\alpha_i^2$$

Under H_{0A} the MSSA is an unbiased estimate of σ_e^2.

8.7.2 Expected Value of Sum of Squares due to Factor B

Proceeding similarly, or by symmetry, we have

$$E(SSB) = E\left[mp\sum_{j=1}^{q}\left(\bar{y}_{.j.} - \bar{y}_{...}\right)^2 \right]$$

Substituting the value of $\bar{y}_{.j.}$ and $\bar{y}_{...}$ from the model

we get $$E(SSB) = E\left[mp\sum_{i=1}^{q}\left(\beta_j + \bar{e}_{.j.} - \bar{e}_{...}\right)^2 \right]$$

or $$E(SSB) = \left[mp\sum_{i=1}^{q}\left\{\beta_j^2 + \left(\bar{e}_{.j.} - \bar{e}_{...}\right)^2 + 2\beta_j\left(\bar{e}_{.j.} - \bar{e}_{...}\right)\right\} \right]$$

or $$E(SSB) = mp\sum_{j=1}^{q}\beta_j^2 + mp\,E\left[\sum_{j=1}^{q}\left(\bar{e}_{.j.} - \bar{e}_{...}\right)^2\right] + 0$$

or $$E(SSB) = mp\sum_{j=1}^{q}\beta_j^2 + mp\,E\left[\sum_{j=1}^{q}\left(\bar{e}_{.j.}\right)^2 - q\,\bar{e}_{...}^2\right]$$

$$= mp\sum_{j=1}^{q}\beta_j^2 + mp\left[\sum_{j=1}^{q}E\left(\bar{e}_{.j.}\right)^2 - qE\left(\bar{e}_{...}^2\right)\right]$$

$$= mp\sum_{j=1}^{q}\beta_j^2 + mp\left[\sum_{j=1}^{q}\frac{\sigma_e^2}{mp} - q\frac{\sigma_e^2}{mpq}\right]$$

$$= mp\sum_{j=1}^{q}\beta_j^2 + q\,\sigma_e^2 - \sigma_e^2$$

$$E(SSB) = (q-1)\sigma_e^2 + mp\sum_{j=1}^{q}\beta_j^2$$

$$E\left(\frac{SSB}{q-1}\right) = \sigma_e^2 + \frac{mp}{(q-1)}\sum_{j=1}^{q}\beta_j^2$$

or $$E(MSSB) = \sigma_e^2 + \frac{mp}{(q-1)}\sum_{j=1}^{q}\beta_j^2$$

Under H_{0B} the MSSB is an unbiased estimate of σ_e^2. Similarly you can obtain the expected value of SSAB , which will be

$$E\ (SSAB) = m\sum_{i=1}^{p}\sum_{j=1}^{q}(\alpha\beta)_{ij}^2 + (p-1)(q-1)\sigma_e^2$$

$$E\left[\frac{SSAB}{(p-1)(q-1)}\right] = \sigma_e^2 + \frac{m}{(p-1)(q-1)}\sum_{i=1}^{p}\sum_{j=1}^{q}(\alpha\beta)_{ij}^2$$

or $$E\ (MSSAB) = \sigma_e^2 + \frac{m}{(p-1)(q-1)}\sum_{i=1}^{p}\sum_{j=1}^{q}(\alpha\beta)_{ij}^2$$

Under H_{0AB}, the mean sum of squares due to interaction between Factor A and B is an unbiased estimate of σ_e^2.

8.7.3 Expected Value of Sum of Squares due to Error

Proceeding similarly, or by symmetry, we have

$$E\,(SSE) = E\left(\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(y_{ijk} - \bar{y}_{ij.}\right)^2\right)$$

Substituting the value of y_{ijk} and $\bar{y}_{ij.}$ from the model, we have

$$E(SSE) = E\left[\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m}\left(e_{ijk} - \bar{e}_{ij.}\right)^2\right]$$

$$= E\left[\sum_{i=1}^{p}\sum_{j=1}^{q}\left(\sum_{k=1}^{m}\left(e_{ijk} - \bar{e}_{ij.}\right)^2\right)\right]$$

$$= E\left[\sum_{i=1}^{p}\sum_{j=1}^{q}\left(\sum_{k=1}^{m} e_{ijk}^2 - m\,\bar{e}_{ij.}^2\right)\right]$$

$$= E\left[\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} e_{ijk}^2 - m\sum_{i=1}^{p}\sum_{j=1}^{q}\bar{e}_{ij.}^2\right]$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{m} E\left(e_{ijk}^2\right) - m\sum_{i=1}^{p}\sum_{j=1}^{q} E\left(\bar{e}_{ij.}^2\right)$$

$$= mpq\ \sigma_e^2 - mpq\,\sigma_e^2/m$$

$$= (mpq - 1)\,\sigma_e^2$$

or $$E\left(\frac{SSE}{pq\,(m-1)}\right) = \sigma_e^2$$

or $$E\,(MSSE) = \sigma_e^2$$

Hence, mean sum of squares due to error is an unbiased estimate of σ_e^2.

Example 1**:** A manufacturer wishes to determine the effectiveness of four types of machines (A, B, C and D) in the production of bolts. To accumulate this, the numbers of defective bolts produced for each of two shifts in the results are shown in the following table:

Machine	First shift					Second Shift				
	M	T	W	Th	F	M	T	W	Th	F
A	6	4	5	5	4	5	7	4	6	8
B	10	8	7	7	9	7	9	12	8	8
C	7	5	6	5	9	9	7	5	4	6
D	8	4	6	5	5	5	7	9	7	10

Perform an analysis of variance to determine at 5% level of significance, whether there is a difference (a) Between the machines and (b) Between the shifts.

Solution: There are two factors the machine and shift. The levels of machine are four and levels of shift are two. The Computation results are as follows:

$$G = 6+4+5+5+4+5+7+4+6+8+10+8+7+7+9+7+9+12+8+8+7+5+6+5+9+9+7+5+4+6+8+4+6+5+5+7+9+7+10$$

$$= 268$$

$$N = 40$$

$$CF = \frac{G^2}{N} = \frac{268 \times 268}{40} = 1795.6$$

Raw Sum of Squares (RSS) = $6^2 + 4^2 + \ldots\ldots + 10^2 = 1946$

Total Sum of Squares (TSS) = RSS - CF = 1946 - 1795.6 = 150.4

Sum of square due to machines and due to shifts can be calculated by considering the following two-way table:

Machine	Shift		Total
	I Shift	II Shift	
A	24	30	54
B	41	44	85
C	32	31	63
D	28	38	66
Total	125	143	268

Sum of Squares due to Machine $(SSM) = \frac{54^2}{10} + \frac{85^2}{10} + \frac{63^2}{10} + \frac{66^2}{10} - CF$

$$= 1846.6 - 1795.6 = 51.0$$

Sum of Squares due to Shifts $(SSS) = \frac{125^2}{20} + \frac{143^2}{20} - CF$

$$= 1803.7 - 1795.6 = 8.1$$

Sum of Squares due to Interaction (SSMS)

$$= \frac{(24)^2}{5} + \frac{(41)^2}{5} + \frac{(32)^2}{5} + \frac{(28)^2}{5} + \frac{(30)^2}{5} + \frac{(44)^2}{5} + \frac{(31)^2}{5} + \frac{(38)^2}{5} - CF - (SSM) - (SSS)$$

$$= 1861.2 - 1795.6 - 51.0 - 8.1 = 6.5$$

Finaly, the Sum of Squares due to error is founded by subtracting the SSM, SSS and SSSM from TSS

$$SSE = TSS - SSM - SSS - SSMS$$

$$= 150.4 - 51.0 - 8.1 - 6.5 = 84.8$$

$$\text{MSSM} = \frac{\text{SSM}}{\text{df}} = \frac{51.0}{3} = 17$$

$$\text{MSSS} = \frac{\text{SSS}}{\text{df}} = \frac{8.1}{1} = 8.1$$

$$\text{MSSE} = \frac{\text{SSE}}{\text{df}} = \frac{84.8}{31} = 2.65$$

$$\text{MSSMS} = \frac{\text{SSMS}}{\text{df}} = \frac{6.5}{3} = 2.167$$

For testing H_{0A} : Mean effect of Machine A= Machine B =

Machine C = Machine D, is

$$F = \frac{17}{2.65} = 6.42$$

For testing H_{0B} : Mean effect of Shift A = Shift B, is

$$F = \frac{8.1}{2.65} = 3.06$$

Similarly, for testing H_{0AB} : Interaction effect of Machine and Shift, is

$$F = \frac{2.167}{2.65} = 0.817$$

ANOVA Table for Two-way Classified Data m- Observation per Cell

Sources of Variation	**Degrees of Freedom (DF)**	**Sum of Squares (SS)**	**Mean Sum of Squares (MSS)**	**F-test or Variance Ratio**
Due to Machinery	3	51.0	17	$\frac{17}{2.65} = 6.42$
Due to Shift	1	8.1	8.1	$\frac{8.1}{2.65} = 3.06$
Due to Interaction	3	6.5	2.167	$\frac{2.167}{2.65} = 0.817$
Due to Error	32	84.8	2.65	
Total	39	150.4		

The tabulated value of F at 3 and 32 degrees of freedom at 5% level of significance is 2.90. The computed value of F for interaction is 0.817 so the average performances in different shifts are not significant. There is a significant difference among machines, since the calculated value of F for machines is 6.42 and the critical value (tabulated value) of F is 2.90. The tabulated value for shifts is 4.15. The calculated value of F for shifts is 3.06. Hence, there is no difference due to shifts.

E1) An experiment is performed to determine the effect of two advertising campaigns on three kinds of cake mixes. Sales of each mix were recorded after the first advertising campaigns and then after the second advertising campaign. This experiment was repeated three times for each advertising campaign and got the following results:

	Campaign I	Campaign II
Mix1	574, 564, 550	1092, 1086, 1065
Mix2	524, 573, 551	1028, 1073, 998
Mix3	576, 540, 592	1066, 1045, 1055

Perform an analysis of variance to determine at 5% level of significance, whether there is a difference (a) Between the cake mixes and (b) Between the campaigns.

8.8 SUMMARY

In this unit, we have discussed:

1. The ANOVA model for two-way classified data with m observations per cell;
2. The basic assumptions for the given model;
3. How to obtain the estimates of the parameters of the given model;
4. How to test the hypothesis for two-way classified data with m observations per cell;
5. How to derive the expectations of the various sum of squares; and
6. Numerical problems to test the hypothesis for two-way classified data with m observations per cell.

8.9 SOLUTIONS /ANSWERS

E1) For set up an ANOVA Table for this problem, the computation results are as follows:

Grand Total G = 14552

N = 18

Correction Factor (CF) = (14552 ×14552) /18 = 11764483.55

RSS = 12882026

TSS = 1117542

SSA = 1107070

SSB = 2957

SSAB = 1126

SSE = 6389

ANOVA Table

Sources of Variation	**DF**	**SS**	**MSS**	**F-Calculated**	**F-Tabulated at 5% level of significance**
Advertising campaign	1	1107070	1107070	1107070/532.42 = 2079.32	F(1,12) = 243.9
Cake Mix	2	2957	1478.5	1478.5/532.42 = 2.8	F(2,12) = 19.41
Interaction	2	1126	563	563/532.42 = 1.06	F(2,12) = 19.41
Error	12	6389	532.42		
Total	17	1117542			

Since computed value of F for cake mix and interaction are 2.8 and 1.06 respectively which are less than corresponding tabulated value so they are not significant. Whereas the calculated value of F for advertising campaign is greater than corresponding tabulated value so there is a significant difference among advertising campaign.

TABLE: The F Table

Value of F Corresponding to 5% (Normal Type) and 1% (Bold Type) of the Area in the Upper Tail

Degrees of Freedom: (Denominator)	Degrees of Freedom (Numerator) 1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	254
	4,052	**4,999**	**5,403**	**5,625**	**5,764**	**5,859**	**5,928**	**5,981**	**6,022**	**6,056**	**6,082**	**6,106**	**6,142**	**6,169**	**6,208**	**6,234**	**6,258**	**6,366**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.50
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.34**	**99.36**	**99.38**	**99.40**	**99.41**	**99.42**	**99.43**	**99.44**	**99.45**	**99.46**	**99.47**	**99.50**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.53
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**	**26.92**	**26.83**	**26.69**	**26.60**	**26.50**	**26.12**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.63
	22.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**	**14.24**	**14.15**	**14.02**	**13.93**	**13.83**	**13.46**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.36
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.27**	**10.15**	**10.05**	**9.96**	**9.89**	**9.77**	**9.68**	**9.55**	**9.47**	**9.38**	**9.02**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.67
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**	**7.60**	**7.52**	**7.39**	**7.31**	**7.23**	**6.88**
7	5.59	4.47	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.23
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**	**6.35**	**6.27**	**6.15**	**6.07**	**5.98**	**5.65**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	2.93
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**	**5.56**	**5.48**	**5.36**	**5.28**	**5.20**	**4.86**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.71
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**	**5.00**	**4.92**	**4.80**	**4.73**	**4.64**	**4.31**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.54
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**	**4.60**	**4.52**	**4.41**	**4.33**	**4.25**	**3.91**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.40
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**	**4.29**	**4.21**	**4.10**	**4.02**	**3.94**	**3.60**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.30
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**	**4.05**	**3.98**	**3.86**	**3.78**	**3.70**	**3.36**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.21
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**	**3.85**	**3.78**	**3.67**	**3.59**	**3.51**	**3.16**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.13
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**	**3.43**	**3.34**	**3.00**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.07
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**	**3.29**	**3.20**	**2.87**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.01
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**	**3.45**	**3.37**	**3.25**	**3.18**	**3.10**	**2.75**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	1.96
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.95**	**3.52**	**3.45**	**3.35**	**3.27**	**3.16**	**3.08**	**3.00**	**2.65**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	1.92
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**	**3.27**	**3.19**	**3.07**	**3.00**	**2.91**	**2.57**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	1.88
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**	**3.19**	**3.12**	**3.00**	**2.92**	**2.84**	**2.49**
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.84
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**	**3.13**	**3.05**	**2.94**	**2.86**	**2.77**	**2.42**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.81
74	**8.02**	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**	**3.07**	**2.99**	**2.88**	**2.80**	**2.72**	**2.36**

TABLE (Continued)

Degrees of Freedom: Denominator	Degrees of Freedom: Numerator																	
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	∞
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.23	2.23	2.18	2.13	2.07	2.03	1.98	1.78
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**	**3.02**	**2.94**	**2.83**	**2.75**	**2.67**	**2.31**
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	3.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.76
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**	**2.97**	**2.89**	**2.78**	**2.70**	**2.62**	**2.26**
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.73
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**	**2.93**	**2.85**	**2.74**	**2.66**	**2.58**	**2.21**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.71
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**	**2.89**	**2.81**	**2.70**	**2.62**	**2.54**	**2.17**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.69
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**	**2.86**	**2.77**	**2.66**	**2.58**	**2.50**	**2.13**
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.67
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**	**2.83**	**2.74**	**2.63**	**2.55**	**2.47**	**2.10**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.65
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**	**2.80**	**2.71**	**2.60**	**2.52**	**2.44**	**2.06**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.64
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**	**2.77**	**2.68**	**2.57**	**2.49**	**2.41**	**2.03**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.62
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**	**2.74**	**2.66**	**2.55**	**2.47**	**2.38**	**2.01**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.00
	6.64	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**	**2.07**	**1.99**	**1.87**	**1.79**	**1.69**	**1.00**

Indira Gandhi National Open University
School of Sciences

MST - 005
Statistical Techniques

Block

3

DESIGN OF EXPERIMENTS

UNIT 9
Completely Randomised Design **5**

UNIT 10
Randomised Block Design **19**

UNIT 11
Latin Square Design **35**

UNIT 12
Factorial Experiments **49**

Curriculum and Course Design Committee

Prof. K. R. Srivathsan
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Parvin Sinclair
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Geeta Kaicker
Director, School of Sciences
IGNOU, New Delhi

Prof. R. M. Pandey
Department of Bio-Statistics
All India Institute of Medical Sciences
New Delhi.

Prof. Jagdish Prasad
Department of Statistics
University of Rajasthan, Jaipur

Prof. Rahul Roy
Maths and Stat. Unit
Indian Statistical Institute, New Delhi

Dr. Diwakar Shukla
Department of Maths and Statistics
Dr. Hari Singh Gaur University, Sagar

Prof. G. N. Singh
Department of Applied Maths
I S M Dhanbad

Prof. Rakesh Srivastava
Department of Statistics
M. S. University of Baroda, Vadodara

Dr. Gulshan Lal Taneja
Department of Mathematics
M. D. University, Rohtak

Faculty Members, School of Sciences, IGNOU

Statistics
Dr. Neha Garg
Dr. Nitin Gupta
Mr. Rajesh Kaliraman
Dr. Manish Trivedi

Mathematics
Dr. Deepika Garg
Prof. Poornima Mital
Prof. Sujatha Varma
Dr. S. Venkataraman

Block Preparation Team

Content Editor
Prof. G. K. Shukla
Decision Science Group
Indian Institute of Management, Lucknow

Language Editor
Dr. Parmod Kumar
School of Humanities, IGNOU

Secretarial Support
Mr. Deepak Singh

Course Writer
Prof. Jagdish Prasad
Department of Statistics
University of Rajasthan, Jaipur

Formated By
Dr. Manish Trivedi
Mr. Prabhat Kumar Sangal
School of Sciences, IGNOU

Programme and Course Coordinator: Dr. Manish Trivedi

Block Production

Mr. Y. N. Sharma, SO (P.)
School of Sciences, IGNOU

Acknowledgement: We gratefully acknowledge Prof. Geeta Kaicker, Director, School of Sciences for her great support and guidance.

December, 2011

ISBN – 978-81-266-5786-5

Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by Director, School of Sciences.

Laser Type set by: Rajshree Computers, V-166A, Bhagwati Vihar, (Near Sector-2, Dwarka), Uttam Nagar, New Delhi-110059

Printed at: Gita Offset Printers Pvt. Ltd., C-90, Okhla Industrial Area, Phase-I, New Delhi-110020.

DESIGN OF EXPERIMENTS

In Block 2, we have discussed the concept of analysis of variances, its definition, terminology used, assumptions of analysis of variance and its uses. The concept of linear models in one-way as well as in two-way analysis of variance technique is also explained. In this block, we shall discuss some experimental designs.

In any field of study either in life sciences or some other it is essential to plan an experiment, i.e. what is the object and which type of data is required. In order to make use of time and energy spent on experiment, it should be planned with a careful designing.

Design of experiments is a logical construction of the experiment in which the degree of uncertainty with which the inference about the population is drawn may be well defined. The experimental designs are formed by following these steps:

1. Planning of the experiments;
2. Obtaining relevant information from it regarding the statistical hypothesis under study;
3. Making a statistical analysis of the data.

Once a design of experiment is decided, the observations are obtained from it and with the technique of analysis of variance, the data is analysed.

In Unit 9 of this block, we have discussed the basic principles of design of experiments. The layout and statistical analysis of the completely randomised design is also discussed in this unit. In Unit 10 we have elaborated the basic idea about the randomised block design with its layout and statistical analysis. Similarly, the layout and statistical analysis of the latin square design is discussed in Unit 11. In the last unit of this block, Unit 12, we have explored the basic idea about the factorial experiments with the layout and statistical analysis of 2^2 and 2^3 factorial experiments.

Suggested Readings:

1. Cochran, W. G. and Cox, G. M.; Experimental Designs (Chs. 1-7, 14), Asia Publishing House, 1959.
2. Federer, W. T.; Experimental Design (Chs. 1-7, 9, 10, 14, 16), Macmillan, 1963, and Oxford & I.B.H., 1967.
3. Fisher, R. A.; The Design of Experiment, Oliver and Boyd, 1947.
4. Kempthorne, O.; The Design and Analysis of Experiments (Chs. 1-3, 5-11, 13-15, 28), John Wiley, 1965, and Wiley Eastern.
5. Mann, H. B.; Analysis and Design of Experiments, Dover, 1949.
6. Quenouille, M. H.; The Design and Analysis of Experiment (Chs. 1-3), Charles Griffin, 1953.
7. Yates, F.; The Design and Analysis of Factorial Experiments (Chs. 1-4, 16), Imperial Bureau of Science, Tech. Com. No. 35, 1937.

Notations and Symbols

y_{ij}	:	The j^{th} observation in the i^{th} level of a factor A
μ `	:	An over all mean or grand mean
y_{ijk}	:	k^{th} observation / response / dependent variable under i^{th} level of factor A and j^{th} level of factor B
μ_i	:	Mean of i^{th} level of a factor
α_i	:	Effect due to i^{th} level of factor A
β_j	:	Effect due to j^{th} level of factor B
$(\alpha\beta)_{ij}$	:	Interaction effect between the i^{th} level of factor A and j^{th} level of factor B
e_{ij}	:	Error term
n_i	:	Number of observations in i^{th} level of factor
E	:	Residual sum of squares
H_0	:	Null hypothesis
H_1	:	Alternative hypothesis
$\bar{y}_{i.}$	:	$\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$
$\bar{y}_{..}$	:	$\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$
N	:	$\sum_{i=1}^{k} n_i$ = Total number of observations
TSS	:	Total Sum of Squares
SST	:	Sum of Squares due to Treatments or different levels of a factor
SSE	:	Sum of Squares due to Error
DF	:	Degrees of freedom
$V(e_{ij})$	:	Variance of error e_{ij}
F	:	F- statistic/ Variation Ratio
α	:	Level of significance
CRD	:	Completely Randomised Design
RBD	:	Randomised Block Design
LSD	:	Latin Square Design
CD	:	Critical Difference
p	:	Number of levels of factor A
q	:	Number of levels of factor B
SSA	:	Sum of Squares due to factor A
SSB	:	Sum of Squares due to factor B

UNIT 9 COMPLETELY RANDOMISED DESIGN

Structure

9.1 Introduction
- Objectives
- Planning of an Experiment
- Classification of Experimental Designs

9.2 Basic Definitions of Experimental Design

9.3 Principles of Design of Experiments
- Randomisation
- Replication
- Local Control

9.4 Size and Shape of the Plots

9.5 Completely Randomised Design
- Layout of Completely Randomised Design
- Statistical Analysis of Completely Randomised Design
- Least Square Estimates of Effects
- Variance of the Estimates
- Expectation of Sum of Squares

9.6 Suitability of CRD

9.7 Summary

9.8 Solutions/Answers

9.1 INTRODUCTION

The modern concepts of experimental designs were primarily given by Ronald A. Fisher in the 1920s and 1930s at "Rothamasted Experimental Station", an agricultural research station of London. In Fisher's first book on design of experiments, he showed how valid conclusions could be drawn efficiently from experiments with natural fluctuation such as temperature, soil conditions and rainfall, that is, in the presence of nuisance variables. The known nuisance variables usually cause systematic biases in groups of results (e.g. batch-to-batch variables). The unknown nuisance variables usually cause random variability in the results and are called inherent variability or noise.

The experimental design was first used in an agricultural context, the method has been applied successfully in the military and in industry since the 1940s. Besse Day, working at U. S. Naval Experimentation Laboratory, used experimental designs to solve problems such as finding the cause of bad welds at the naval shipyards during World War II. George Box, employed by Imperial Chemical Industries before coming to the United States, is a leading developer of experimental design produced for optimizing chemical process. W. Edwards Deming taught statistical methods, including experimental designs, to Japanese scientist and engineers in the early 1950's at a time when "Made in Japan" meant poor quality. Genichi Taguchi, the most well known of this group of Japanese scientists is famous for his quality improvement methods. One of the companies where Taguchi first applied his methods was Toyota. Since the late 1970's, U.S. industry has become interested again in

quality improvement initiatives, now known as “Total Quality” & “Six-sigma” programs. Design of experiments is considered an advanced method in the six sigma programs, which were pioneered at Motorola & GE.

According to Bernad Ostle, “The design of experiment is, the 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 to valid inferences with respect to stated problem”.

In any field of study either in life sciences or some other, it is essential to plan an experiment, i.e. what is the object and which type of data is required. In order to make use of time and energy spent on experiment, it should be planned with a careful designing. Once a design of experiment is decided, the observations are obtained from it and with the technique of analysis of variance, the data is analysed.

Basic definitions of experimental design are described in Section 9.2. The principles of design of experiments i.e. randomisation, replication and local control are explained in Section 9.3 whereas size and shape of plots are described in Section 9.4. In Section 9.5, layout and statistical analysis of completely randomised design are explained. The least square estimates of effects, variance of the estimates and expectation of sum of squares are also given in Section 9.5. Suitability of CRD is descrived in Section 9.6.

Objectives

After studying this unit, you would be able to

- describe the experimental design;
- explain the planning and classification of experimental designs;
- describe the principles of design of experiments;
- explain the completely randomised design;
- describe the layout of CRD;
- explain the statistical analysis of CRD; and
- explain the advantages and disadvantages as well as the suitability of CRD.

9.1.1 Planning of an Experiment

There are some basic points regarding the planning of an experiment, which should be under consideration. These are as follows:

1. **The Experiment should be Free from Bias**

   An experiment must be planned so that it gives an unbiased estimate of the values we wish to measure. It is a matter of the design being such that no bias on the part of the experimenter can possibly enter into the results. This is achieved mainly by randomisation.

2. **There must be a Measure of Error**

   The true experiment is one that is strictly objective. It should furnish a measure of error and this error alone should be the measuring stick of significance.

3. **There must be a Clearly Defined Objective**

   For an experiment it is essential to specify the objects perfectly. In other words the objective of the experiment should be clearly defined.

4. **The Experiment should have Sufficient Accuracy**
   The accuracy of an experiment can be brought by the elimination of technical errors and by increasing replications. The number of replications should be decided to produce a given degree of accuracy.

9.1.2 Classification of Experimental Designs

Statisticians by themselves do not design experiments, but they have developed a number of structured schedules called "experimental designs", which they recommend for the taking of measurements. These designs have certain rational relationships to the purposes, needs and physical limitations of experiments. Designs also offer certain advantages in economy of experimentation and provide straightforward estimates of experimental effects and valid estimates of variance. There are a number of ways in which experiment designs might be classified, for example, the following:

1. By the number of experimental factors to be investigated (e.g., single-factor versus multifactor designs)
2. By the structure of the experimental design (e.g., blocked, factorial, nested, or response-surface design)
3. By the kind of information which the experiment is primarily intended to provide (e.g. estimates of effects, estimates of variance, or empirical mappings).

9.2 BASIC DEFINITIONS OF EXPERIMENTAL DESIGN

Several fundamental terms are widely used throughout this section. They may be defined as follows:

1. **Treatment**
   In an experiment, there are some variants under study, the effects of which are measured and tested (compared). These variants will be referred to as treatments. For example, to test the effects of three fertilizers, i.e., Nitrogen, Phosphorus and Potash on the yield of a certain crop. Then Nitrogen, Phosphorus and Potash are called treatments.

2. **Yield**
   The response of the treatment is measured by some indicator such as crop production, milk production, body temperature, mileage of engine set, etc. Such an indicator is called yield. The treatments are applied to some units such as field plots, sample of cows, sample of patients, sample of engine, sets, etc. and the effect on the yield is observed.

3. **Experimental Units**
   A unit to which one treatment applied is called experimental unit. It is the smallest division of an experimental material to which the treatment applied and on which the variable under study is measured. In carrying out an experiment, we should clear as to what constitute the experimental unit.

It can be understood that in a field of agriculture it is called plot, in the field of animal husbandry it may be a cow (cattle), in the field of medicine it may be a patient and in the field of automobile industry it may be engine set and so on.

4. **Experimental Material**
We have already explained the concept of experimental unit. The experimental material is nothing but a set of experimental units. For example, a piece of land, a group of cows, a number of patients and a group of engine sets, etc. Actually, an experimental material is that material on which some set of treatments are applied and tested.

5. **Blocks**
The experimental material is divided into a number of groups or strata which are so formed that they are within homogeneous and between heterogeneous. These groups or strata are called blocks.

6. **Experimental Error**
There is always a variation between the yields of the different plots even when they get the same treatment. This variation exists due to non-assignable causes, which cannot be detected and explained. These are taken to be of random type. This unexplained random part of variation is termed as experimental error. This include all types of extraneous variation due to, (i) inherent variability in the experimental units, (ii) error associated with the measurement made and (iii) lack of representativeness of the sample of the population understudy.

7. **Precision**
The precision of an experiment is measured by the reciprocal of the variance of a mean, i.e.

$$\frac{1}{v(\bar{x})} = \frac{1}{\sigma^2_{\bar{x}}} = \frac{n}{\sigma^2}$$

As n, the replication number increases, precision also increases.

8. **Uniformity Trial**
We know that to increase the efficiency of a design, the plots should be arranged into homogeneous blocks. It can be done only if we have a correct idea about the fertility variation of the field. This is achieved through uniformity trial. It is known that fertility of soil does not increase or decrease uniformly in any direction but it is distributed over the entire field in an erratic manner. By a uniformity trial, we mean a trial in which the field (experimental material) is divided into small units (plots) and the same treatment is applied on each of the units and their yields are recorded. From these yields we can draw a fertility control map which gives us a graphic picture of the variation of the soil fertility and enables us to form a good idea about the nature of the soil fertility variation. This fertility control map is obtained by joining the points of equal fertility through lines.
A uniformity trial gives us an idea about the

1. Fertility gradient of the field,
2. Determination of the shape of the plots to be used,
3. Optimum size of plots,
4. Estimation of number of replications required for achieving certain degree of accuracy.

9.3 PRINCIPLES OF DESIGN OF EXPERIMENTS

Good experimentation is an art and depends heavily upon the prior knowledge and abilities of the experimenter. Designing an experiment means deciding how the observations or measurements should be taken to answer a particular question in a valid, efficient and economical way. If a design is properly designed, then there will exists an appropriate way of analsing the data. From an ill-designed experiment no conclusion can be drawn.

The fundamental principles in design of experiments are the solutions to the problems in experimentation posed by the two types of nuisance factors and serve to improve the efficiency of experiments. For the validity of the design Prof. R.A. Fisher gave three principles of design of experiments, those fundamental principles are:

- Randomisation
- Replication
- Local Control

9.3.1 Randomisation

The principle of randomisation is essential for a valid estimate of the experimental error and to minimize the bias in the results. In the words of Cochran and Cox, “Randomisation is analogous to insurance in that it is a precaution against disturbances that may or may not occur and they may or may not be serious if they do occur”. Thus, randomisation is so done that each treatment should get an equal chance. We mean that the treatments should be allocated randomly, i.e., by the help of random numbers. The following are the advantages of randomisations:

1. It provides a basis for the test of significance because randomisation ensures the independence of the observations which is one of the assumptions for the analysis of variance.

2. It is also a device for eliminating bias. Bias creeps in experiment, when the treatments are not assigned randomly to the units. This bias may be personal or subjective. The randomisation ensures the validity of the results.

9.3.2 Replication

“Replication” is the repetition, the rerunning of an experiment or measurement in order to increase precision or to provide the means for measuring precision. A single replicate consists of a single observation or experimental run. Replication provides an opportunity for the effects of uncontrolled factors or factors unknown to the experimenter to balance out and thus, through randomisation, acts as a bias-decreasing tool. Suppose a pain relieving drug A is applied to 4 patients, we say that drug A is replicated four times. By repeating a treatment it is possible to obtain a more reliable estimate because it reduces the experimental error. Further by repeating a treatment number of times we can judge the average performance of a treatment and the situation becomes clearer. Basically there are following uses of replication:

1. It enables us to obtain a more precise estimate of the treatments effects.
2. The next important purpose of replication is to provide an estimate of the experimental error without which we cannot test the significance of the difference between any two treatments. The estimate of experimental error is obtained by considering the difference in the plots receiving the same treatment in different replications and there is no other alternative of obtaining this estimate.
3. For a desired amount of precision, the minimum number of replications can be obtained.

9.3.3 Local Control

This method is used to attain the accuracy or to reduce the experimental error without increasing unduly the number of replications. Local control is a technique that handles the experimental material in such a way that the effects of variability are reduced. In local control, experimental units are divided into a number of homogeneous groups called blocks. These blocks are so formed that they are homogeneous within and heterogeneous between. This blocking of experiment may be row-wise, column-wise or both according to the number of factors responsible for heterogeneity. Different types of blocking constitute different types of experimental designs. The following are the advantages of local control:

1. By means of local control, the experimental error is reduced considerably and the efficiency of the design is increased.
2. By means of local control the test procedure becomes more sensitive or powerful.

Besides the above three principles, there are some other general principles in designing an experiment. Familiarity with the treatments and experimental material is an asset. Selection of experimental site is an asset. Selection of experimental site should be carefully done. Within block variability should be reduced.

9.4 SIZE AND SHAPE OF THE PLOTS

In field experiments, the size and shape of plots as well as of blocks influence the experimental error. The total available experimental area remaining fixed, an increase in size of plots will automatically decrease the number of plots and indirectly increases the block size. In order to reduce the flow of experimental material from one plot to another, it is customary to leave out strips of land between consecutive plots and also between blocks. These non-experimental areas are known as guard area. The size and shape of the plot should be such that we make a compromise between statistical and practical requirements i.e. if plot size is x and the variance of the plot is V(x), then V(x) is minimum (statistical consideration) and there should be no disturbance for agricultural operations (practical requirements).

The size and shape of block will ordinarily be determined by the size and shape of plots and the number of plots in a block. It is desirable from the point of view of error control to have small variations among the plots within a block and large variation among the blocks i.e. in general the division of experimental material into blocks is made in such a way that plots within blocks are as homogeneous as possible.

Based on these fundamental principles, we have certain designs. The analysis of those designs is based on the theory of least squares which gives the best estimates of the treatments effects and was initiated by Fisher (1926) followed by Yates (1936), Bose & Nair (1939) and Rao (1976). The following three designs are frequently used:

1. Completely Randomised Design
2. Randomised Block Design
3. Latin Square Design

9.5 COMPLETELY RANDOMISED DESIGN

The simplest of all the design is completely randomised design (CRD) which is applied in the case when the experimental materials are homogeneous. CRD is based on two principles i.e. randomisation and replication. The third principle, i.e. local control is not used because it is assumed that experimental materials are homogeneous. In this, the treatments are allocated randomly to the experimental units and each treatment is assigned to different experimental units completely at random (can be repeated any number of times) that is why it is called completely randomised design.

Suppose we have k treatments under comparison and the i^{th} treatment is to be replicated n_i times for i =1, 2, …, k, then the total number of units required for the design are $n = \sum_{i=1}^{k} n_i$. We allocate the k treatments completely at random to n units such that i^{th} treatment appears n_i times in the experiment.

9.5.1 Layout of Completely Randomised Design

The term layout refers to the allocation of different treatment to the experimental units. We have already said that treatments are allocated completely at random to the different experimental units. Every experimental unit has the same chance of receiving a particular treatment.

Suppose we want to test the effect of three pain relieving drugs A, B and C on twelve patients. Then we first number all the patients (units) from 1 to 12. Then from a random number table of one digit we pick up 12 numbers which are less than 4. Suppose the numbers are 1, 3, 2, 1, 3, 2, 1, 3, 2, 2, 3, 1. Thus the drug A is allotted to patient 1, drug C is allotted to patient number 2 and so on. It can be shown below:

(1) A	(2) C	(3) B	(4) A	(5) C	(6) B
(7) A	(8) C	(9) B	(10) B	(11) C	(12) A

It is clear from the above layout that the replications of A, B and C are equal. If the number of replications for each treatment is 5, 4 and 3 respectively, we number the experimental units in a convenient way from 1 to 12. We then get a random permutation of the experimental units. To the first 5 of the units in the random permutation we assign treatment A, to the next 4 units treatment B is assigned and the treatment C is assigned to the remaining 3 units.

9.5.2 Statistical Analysis of Completely Randomised Design

Statistical analysis of a CRD is analogous to the ANOVA for one-way classified data for fixed effect model, the linear model (assuming various effect to be additive) becomes

$$y_{ij} = \mu + \alpha_i + e_{ij}\ ,\ i = 1, 2, 3, \ldots, k;\ \ j = 1, 2, 3, ..., n_i \qquad \ldots(1)$$

where y_{ij} is the yield or response from the j^{th} unit receiving the i^{th} treatment, μ is the general mean effect, α_i is the effect due to the i^{th} treatment, where μ and α_i are constants so that $\sum_{i=1}^{k} n_i \alpha_i = 0$ and e_{ij} is identically and independently distributed (i.i.d.) $N(0, \sigma_e^2)$. Then, $n = \sum_{i=1}^{k} n_i$ is the total number of experimental units.

The analysis of model given in equation (1) is as same as that of fixed effect model of one-way classified data, discussed in Unit 6 of MST-005. If we write

$\sum_i \sum_j y_{ij} = y_{..} = G$ = Grand total of the n observations, and

$\sum_{j=1}^{n_i} y_{ij} = y_{i.} = T_{i.}$ = Total response in the units receiving the i^{th} treatment,

Then, as in ANOVA (one-way classified data),

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} + \bar{y}_{i.}\right)^2 + \sum_{i=1}^{k} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$$

i.e. $\qquad TSS = SSE + SST$

where, TSS, SST and SSE are the Total Sum of Squares, Sum of Squares due to Treatments (between treatments SS) and Sum of Square due to Error (within treatment SS) given respectively by

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2$$

$$SST = \sum_{i=1}^{k} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 = S_T^2$$

and $\quad SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2 = S_E^2$

ANOVA Table for CRD

Source of Variation	DF	SS	MSS	Variance Ratio (F)
Treatments	$k-1$	$SST = S_T^2$	$MSST = \frac{S_T^2}{(k-1)}$	$F_T = \frac{MSST}{MSSE}$
Error	$n-k$	$SSE = S_E^2$	$MSSE = \frac{S_E^2}{(n-k)}$	
Total	$n-1$	$S_T^2 + S_E^2$		

Under the null hypothesis, H_0: $\alpha_1 = \alpha_2 = \ldots. = \alpha_k$ against the alternative that all α's are not equal, the test statistic

$$F_T = \frac{MSST}{MSSE} \sim F(k-1, n-k)$$

i.e., F_T follows F distribution with $(k-1, n-k)$ df.

If $F_T > F_{(k-1, n-k)}\ (\alpha)$ then H_0 is rejected at α level of significance and we conclude that treatments differ significantly. If $F_T < F_{(k-1, n-k)}\ (\alpha)$ then H_0 may be accepted i.e. the data do not provide any evidence to prefer one treatment to the other and as such all of them can be considered alike.

If the treatments show significant effect then we would be interested to find out which pair of treatments differs significantly. For this instead of calculating Student's t-test for different pairs of treatment means we calculate the least significant difference at the given level of significance. This least difference is called as critical different (CD) and CD at α level of significance is given by

CD = Standard error of difference between two treatment means × $t_{\alpha/2}$ for error degrees of freedom.

We have

$$\text{Var}\,(\bar{y}_{i.} - \bar{y}_{.j}) = \frac{\sigma_e^2}{n_i} + \frac{\sigma_e^2}{n_j} = \sigma_e^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)$$

$$\text{Standard Error}\,(\bar{y}_{i.} - \bar{y}_{.j}) = \sigma_e\left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}$$

Hence, the critical difference (CD) for $(\bar{y}_{i.} - \bar{y}_{.j})$

$$= t_{\alpha/2}\,(\text{for error df}) \times \left[MSSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)\right]^{1/2}$$

Since MSSE provides an unbiased estimate of σ_e^2.

If each treatment is replicated n times, that is $n_i = n$ for i=1, 2, ..., k then

$$\text{CD for difference of mean} = (t_{\alpha/2} \text{ for error df}) \times \left[MSSE \times \left(\frac{2}{n}\right)\right]^{1/2}$$

9.5.3 Least Square Estimates of Effects

The completely randomised model in equation (1) in Sub-section 9.5.2 is a fixed effect model. Proceeding exactly as in Section 6.4 of Unit 6, we shall get

$$\hat{\mu} = \frac{y_{..}}{n} = \bar{y}_{..} \text{ and } \hat{\alpha}_i = \bar{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..} \qquad \ldots (2)$$

9.5.4 Variance of the Estimates

Proceeding exactly as in Section 6.7 of Unit 6, we shall get

$$\text{Var}\,(\hat{\mu}) = \frac{\sigma_e^2}{n}; \text{ where } n = \sum_{i=1}^{k} n_i \qquad \ldots (3)$$

and $\quad \text{Var}(\alpha_i) = \text{Var}(\alpha_i) = \sigma_e^2 \left(\frac{1}{n_i} - \frac{1}{\sum_{i=1}^{k} n_i} \right)$ … (4)

If we assume that each treatment is replicated an equal number of times i.e., if $n_i = n$, (say), i = 1, 2, …, k; then $n = \sum_{i=1}^{k} n_i = nk$

Hence, from equations (3) and (4), we get

$$\text{Var}(\hat{\mu}) = \frac{\sigma_e^2}{nk} \text{ and } \text{Var}(\hat{\alpha}_i) = \text{Var}(\alpha_i) = \sigma_e^2 \left(\frac{k-1}{nk} \right) \quad \text{… (5)}$$

9.5.5 Expectation of Sum of Squares

Proceeding exactly as in Section 6.7 of Unit 6 [fixed effect model for one-way classified data], we get

$$E(SST) = E\left[\sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2\right] = (k-1)\ \sigma_e^2 + \sum_{i=1}^{k} n_i \alpha_i^2$$

$$E(MSST) = E\left[\frac{S_T^2}{(k-1)}\right] = \sigma_e^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i \alpha_i^2 \quad \text{… (13)}$$

$$E(SSE) = (n-k)\,\sigma_e^2$$

$$E(MSSE) = E\left[\frac{S_E^2}{(n-k)}\right] = \sigma_e^2 \quad \text{… (14)}$$

The method of analysis of completely randomised design would be similar to one-way ANOVA, which has been illustrated below with the following example:

Example 1: A person wanted to purchase a lot of electric drills. He got quotations from five manufacturers. For the selection, he wanted to conduct an experiment to estimate the time taken by each making a hole in a metallic sheet. As the sheet might not be uniform all over in respect of thickness and hardness, he marked 20 places on the sheet and applied four drills from each concern in 4 randomly selected places to make holes. The time for making each hole was recorded and these formed the observations. The observations in seconds are shown below in brackets along with marks of the drills denoted by D_1, D_2, D_3, D_4 and D_5.

$$D_1(19)\ D_3(22)\ D_4(20)\ D_1(20)$$
$$D_5(29)\ D_2(24)\ D_5(30)\ D_3(24)$$
$$D_2(26)\ D_4(25)\ D_1(16)\ D_2(22)$$
$$D_5(28)\ D_3(25)\ D_5(31)\ D_4(28)$$
$$D_4(27)\ D_1(16)\ D_2(27)\ D_3(20)$$

Conduct the experiment by adopting a completely randomised design.

Solution: The analysis of the given design is done by one-way analysis of variance method. The data is analysed and computation results are given as below:

The totals of time records for 4 holes by each of the different makes are denoted by T_1, T_2, T_3, T_4 and T_5 are shown below.

$$T_1 = 71,\ T_2 = 99,\ T_3 = 91,\ T_4 = 100,\ T_5 = 118$$

Grand Total (G) = 479

$$\text{Correction Factor (CF)} = \frac{G^2}{N} = \frac{(479)^2}{20} = 11472.05$$

$$\text{Total Sum of Squares (TSS)} = 21^2 + 18^2 + 22^2 + \ldots + 31^2 + 20^2 - 11472.05$$

$$= 11847 - 11472.05 = 374.95$$

Sum of Squares due to Makes (SSM)

$$= \frac{(71)^2 + (99)^2 + (91)^2 + (100)^2 + (118)^2}{4} - 11472.05$$

$$= 11761.75 - 11472.05 = 289.70$$

$$\text{Sum of Squares due to Error (SSE)} = \text{TSS} - \text{SSM}$$

$$= 374.95 - 289.70 = 85.25$$

Analysis of Variance Table

Sources of Variation	DF	SS	MSS	F
Makes	4	289.7	72.425	12.75
Error	15	85.25	5.68	
Total	19	374.95		

The tabulated value of F at 1 per cent level of significance for 4 and 15 df is 4.89. Thus, the calculated value of F viz. 12.75 shows that Make to Make variation is highly significant thereby indicating that the hypothesis that the time periods taken by the different Makes in boring a hole are, on an average, the same, is rejected. So multiple camparison test will be applied for different Makes.

Mean for Different Makes

Makes				
$\bar{D}_1$	$\bar{D}_2$	$\bar{D}_3$	$\bar{D}_4$	$\bar{D}_5$
17.74	24.75	22.75	25.00	29.50

$$\text{SE} = \sqrt{\frac{2\text{MSSE}}{n}} = \sqrt{\frac{2 \times 5.68}{4}} = 1.69$$

Critical difference at 1 % level of significance

$$\text{CD} = t_{\alpha/2} \text{ (for error df)} \times \text{SE} = 3.055 \times 1.69 = 5.16$$

The initial difference indicates that the Make D_5 is significantly better than all the other Makes.

Pair of Treatments	Difference	CD	Inference
D_1, D_2	$\lvert\overline{D_1} - \overline{D_2}\rvert = 7.01$	5.16	Significant
D_1, D_3	$\lvert\overline{D_1} - \overline{D_3}\rvert = 5.01$	5.16	Insignificant
D_1, D_4	$\lvert\overline{D_1} - \overline{D_4}\rvert = 7.26$	5.16	Significant
D_1, D_5	$\lvert\overline{D_1} - \overline{D_5}\rvert = 11.26$	5.16	Significant
D_2, D_3	$\lvert\overline{D_2} - \overline{D_3}\rvert = 2.00$	5.16	Insignificant
D_2, D_4	$\lvert\overline{D_2} - \overline{D_4}\rvert = 0.25$	5.16	Insignificant
D_2, D_5	$\lvert\overline{D_2} - \overline{D_5}\rvert = 4.75$	5.16	Insignificant
D_3, D_4	$\lvert\overline{D_3} - \overline{D_4}\rvert = 2.25$	5.16	Insignificant
D_3,D_5	$\lvert\overline{D_3} - \overline{D_5}\rvert = 6.75$	5.16	Significant
D_4, D_5	$\lvert\overline{D_4} - \overline{D_5}\rvert = 4.5$	5.16	Insignificant

E1) Carryout the ANOVA for the given following data of yields of 5 varieties, 7 observations on each variety:

Variety	Observations						
	1	**2**	**3**	**4**	**5**	**6**	**7**
1	13	15	14	14	17	15	16
2	11	11	10	10	15	9	12
3	10	13	12	15	14	13	13
4	16	18	13	17	19	14	15
5	12	12	11	10	12	10	10

9.6 SUITABILITY OF CRD

The following are some situations, in which one can apply the complete randomised design:

1. The CRD is used in the situations where experimental materials are homogeneous. That is why, CRD is mostly used in chemical, biological and banking experiments, where the experimental material is thoroughly mixed powder, liquid or chemical.
2. The CRD is used in the situations where the observations on some units are missing or destroyed. This feature of missing observation does not disturb the analysis of the design.

3. In agricultural experiments, this design is not used because experimental material is not homogeneous.

9.6.1 Advantages and Disadvantages of CRD

Advantages of CRD

1. In this design any number of treatments and replications can be used. There may be different number of replications for different treatments.
2. Analysis is simple and easy even if the number of replication is unequal for each treatment. In such case experimental error will differ from treatment to treatment.
3. If some of the observations are missing or destroyed or not available due to some reasons, the analysis can be done without any problem.
4. It provides large degree of freedom for error sum of squares. This increases the sensitivity of the experiment.
5. In CRD there is no condition on the number of replication of the treatments, they can be increased or decreased according to the need of the experimenter. Thus, the design is flexible.

Disadvantages of CRD

1. The main disadvantage of CRD is that the principle of local control has not been used in this design. Due to this fact, the experimental error is inflated. This is the main reason for the criticism of CRD.
2. In agricultural experiments, the design is seldom used because the experimental material is not homogenous.

9.7 SUMMARY

In this unit, we have discussed:

1. The experimental design;
2. The planning and classification of experimental designs;
3. Principles of design of experiments;
4. Completely randomised design;
5. Layout of CRD;
6. The statistical analysis of CRD; and
7. Advantages and disadvantages as well as suitability of CRD.

9.8 SOLUTIONS / ANSWERS

E1) The analysis of the given design is done by one-way analysis of variance method. The data is analysed and computation results are given as below:

Correction Factor (CF) = 6072.03

Raw Sum of Squares (RSS) = 6293

Total Sum of Squares (TSS) = 220.97

Sum of Squares due to Variety (SSV) = 138.40

Sum of Squares due to Error (SSE) = TSS – SSV

= 220.97 - 138.40 = 82.57

ANOVA Table

Source of Variation	DF	SS	MSS	Variance Ratio	
				Calculated	Tabulated
Variety	4	138.40	34.60	12.58	2.66
Error	30	82.57	2.75		
Total	34	220.97			

Null Hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_5$

Since, calculated value of F is greater than the tabulated value of F, we reject the null hypothesis and conclude that variety effects are significantly different.

Mean for Different Varieties

Varieties				
$\bar{D}_1$	$\bar{D}_2$	$\bar{D}_3$	$\bar{D}_4$	$\bar{D}_5$
14.86	11.14	12.86	16.00	11.00

$$SE = \sqrt{\frac{2MSSE}{n}} = \sqrt{\frac{2 \times 13.34}{7}} = 1.95$$

Critical difference at 1 % level of significance

$= t_{\alpha/2}$ (for error df) $\times$ SE = 3.055 $\times$ 1.95 = 5.96

The initial difference indicates that the Variety D_4 is significantly better than all the other Varieties.

Pair of Treatments	Difference	CD	Inference
D_1, D_2	$\lvert\overline{D_1} - \overline{D_2}\rvert = 3.72$	5.96	Insignificant
D_1, D_3	$\lvert\overline{D_1} - \overline{D_3}\rvert = 2.00$	5.96	Insignificant
D_1, D_4	$\lvert\overline{D_1} - \overline{D_4}\rvert = 1.14$	5.96	Insignificant
D_1, D_5	$\lvert\overline{D_1} - \overline{D_5}\rvert = 3.86$	5.96	Insignificant
D_2, D_3	$\lvert\overline{D_2} - \overline{D_3}\rvert = 1.72$	5.96	Insignificant
D_2, D_4	$\lvert\overline{D_2} - \overline{D_4}\rvert = 4.86$	5.96	Insignificant
D_2, D_5	$\lvert\overline{D_2} - \overline{D_5}\rvert = 0.14$	5.96	Insignificant
D_3, D_4	$\lvert\overline{D_3} - \overline{D_4}\rvert = 3.14$	5.96	Insignificant
D_3, D_5	$\lvert\overline{D_3} - \overline{D_5}\rvert = 1.86$	5.96	Insignificant
D_4, D_5	$\lvert\overline{D_4} - \overline{D_5}\rvert = 5.00$	5.96	Insignificant

UNIT 10 RANDOMISED BLOCK DESIGN

Structure

10.1 Introduction
Objectives
10.2 Layout of Randomised Block Design
10.3 Statistical Analysis of RBD
Least Square Estimates of Effects
Variance of the Estimates
Expectation of Sum of Squares
10.4 Missing Plots Technique in RBD
One Missing Plot
Two Missing Plots
10.5 Suitability of RBD
10.6 Summary
10.7 Solutions /Answers

10.1 INTRODUCTION

The completely randomised design was simple due to the reason that principle of local control was not used and it was assumed that the experimental material is homogeneous, but it is observed that the experimental material is not fully homogeneous. In agricultural field experiments sometimes a fertility gradient is present in one direction. In such situation the simple method of controlling variability of the experimental material consist in stratifying or grouping the whole experimental area into relatively homogeneous strata or sub-groups (called blocks), perpendicular to the direction of fertility gradient. These blocks are so formed that plots within a block are homogeneous and between blocks are heterogeneous. In other words, there may be less variation within a block and major difference or variation between blocks. It is to be kept in mind that familiarity with the nature of experimental units is necessary for an effective blocking of the material. The procedure of division of experimental material into a number of blocks give rise to a design known as Randomised block design (RBD) which can be defined as an arrangement of t treatments in r blocks such that each treatment occurs precisely once in each block.

In other words, when the experimental units are heterogeneous, a part of the variability can be accounted for by grouping the experimental units in such a way that those experimental units within each group are as homogeneous as possible. The treatments are then allotted randomly to the experimental units within each group (or block). This results in an increase in precision of estimates of the treatment contrasts, due to the fact that error variance that is a function of comparisons within blocks is smaller because of homogeneous blocks.

Layout and statistical analysis of randomised block design are explained in Sections 10.2 and 10.3. The least square estimates of effects, variance of the estimates and expectation of sum of squares are also given in Section 10.3.

Missing plots techniques in RBD for one and two missing plots are described in Section 10.4 whereas the suitability of RBD is explored in Section 10.5.

Objectives

After studying this unit, you would be able to

- explain the randomised block design;
- describe the layout of RBD;
- explain the statistical analysis of RBD;
- find out the missing plots in RBD; and
- explain the advantages and disadvantages as well as the suitability of RBD.

10.2 LAYOUT OF RANDOMISED BLOCK DESIGN

The entire experimental material is divided into a number of blocks equal to the number of replications for each treatment. Then each block is divided into a number of plots equal to the number of treatments. For example if we have 4 treatments A, B, C and D and each treatment is to be replicated 3 times. Then according to the condition of RBD, we will arrange the experimental material in three blocks each of size 4, i.e. each block consists of 4 plots. After arranging the experimental material into a number of blocks, treatments are allocated to each block separately. That is randomisation is applied afresh for each block and thus, it will be independent for each block. The method is illustrated below by the following arrangement of 3 blocks and 4 treatments:

Layout of RBD with 4 treatments

Block I	A	B	D	C
Block II	C	A	D	B
Block III	D	B	C	A

10.3 STATISTICAL ANALYSIS OF RBD

If in RBD a single observation is made on each of the experimental units, then its analysis is analogous to ANOVA for fixed effect model for a two-way classified data with one observation per cell and the linear model effects to be (additive) becomes

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} ; \qquad i = 1, 2, \ldots, p; j = 1, 2, \ldots, q.$$

where, y_{ij} is the yield or response of the experimental unit receiving the i^{th} treatment in the j^{th} block, μ is the general mean effect, α_i is the effect due to the i^{th} treatment, β_j is the effect due to j^{th} block or replicate and e_{ij} is identically and independently distributed i.e. e_{ij} follows (i.i.d.) N $(0, \sigma_e^2)$, where μ, α_i and β_j are constants so that $\sum_{i=1}^{p} \alpha_i = 0$ and $\sum_{j=1}^{q} \beta_j = 0$.

If we write that

$$\sum_i \sum_j y_{ij} = y_{..} = G = \text{Grand total of all the } p \times q \text{ observations.}$$

$$\sum_j y_{ij} = y_{i.} = \alpha_i = \text{Total for } i^{th} \text{ treatment}$$

$$\sum_i y_{ij} = y_{.j} = \beta_j = \text{Total for } j^{th} \text{ block}$$

Then heuristically, we get

$$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = \sum_i \sum_j [(\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})]^2$$

$$= q\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 + p\sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

The product terms vanish since the algebraic sum of deviations from mean is zero. Thus

$$TSS = SSE + SSB + SST$$

where TSS, SST, SSB and SSE are the total sum of squares, sum of squares due to treatments (between treatments SS), sum of squares due to blocks and sum of squares due to error (i.e., within treatment SS) given respectively by

$$TSS = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$$

$$SST = q\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 = S_T^2 \text{ (say)}$$

$$SSB = p\sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 = S_B^2$$

$$SSE = S_E^2 = TSS - SSB - SST$$

Hence, the total sum of squares is partitioned into three sum of squares whose degree of freedom make the total to the degree of freedom of TSS.

ANOVA Table for RBD

Source of Variation	DF	SS	MSS	Variance Ratio(F)
Treatments	$p-1$	S_T^2	MSST= $S_T^2/(p-1)$	$F_T = \frac{MSST}{MSSE}$
Blocks	$q-1$	S_B^2	MSSB = $S_B^2/(q-1)$	
Error	$(p-1)(q-1)$	S_E^2	MSSE = $S_E^2/(p-1)(q-1)$	$F_B = \frac{MSSB}{MSSE}$
Total	$pq-1$			

Under the null hypothesis, H_0: $\alpha_1 = \alpha_2 = \ldots. = \alpha_p$ against the alternative that all α's are not equal, the test statistic

$$F_T = \frac{MSST}{MSSE} \quad \text{follows F } [(p-1), (p-1)(q-1)]$$

i.e., F_T follows F-distribution with $[(p-1), (p-1)(q-1)]$ df.

If $F_T \geq F$ with $[(p-1), (p-1)(q-1)]$ df at α level of significance, (Usually 5%) then H_0 is rejected and we conclude that treatments differ significantly.

If $F_T < F$ with $[(p-1), (p-1)(q-1)]$ df at α level of significance then H_0 may be accepted, i.e. the data do not provide any evidence against the null hypothesis which may be accepted.

Similarly, under the null hypothesis, H_0: $\beta_1 = \beta_2 = \ldots. = \beta_q$ against the alternative that all β's are not equal, the test statistic

$$F_T = \frac{MSSB}{MSSE} \quad \text{follows F}[(q-1), (p-1)(q-1)]$$

and we can discuss its significance as explained above.

10.3.1 Least Square Estimates of Effects

Proceeding exactly similar as in CRD, and replacing k by p, n by q and taking N = pq, the estimates of the parameters μ, α_i and β_j are given by:

$$\hat{\mu} = \bar{y}_{..}, \ \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}, \ \hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..} \qquad \ldots (1)$$

10.3.2 Variance of the Estimates

Proceeding exactly similar as in CRD, we shall get

$$\text{Var}(\hat{\mu}) = \frac{\sigma_e^2}{pq}$$

$$\text{Var}(\hat{\alpha}_i) = \frac{(p-1)}{pq}\sigma_e^2$$

and $$\text{Var}(\hat{\beta}_j) = \frac{(q-1)}{pq}\sigma_e^2 \quad .$$

10.3.3 Expectation of Sum of Squares

Proceeding exactly as in CRD, we get

$$E[SST] = (p-1)\sigma_e^2 + q\sum_i \alpha_i^2$$

$$E\left[\frac{(SST)}{(p-1)}\right] = E(MSST) = \sigma_e^2 + \frac{q}{(p-1)}\sum_i \alpha_i^2$$

$$E(SSB) = (q-1)\sigma_e^2 + p\sum_j \beta_j^2$$

$$E\left[\frac{(SSB)}{(q-1)}\right] = E(MSSB) = \sigma_e^2 + \frac{p}{(q-1)}\sum_j \beta_j^2$$

$$E(SSE) = (q-1)(p-1)\sigma_e^2$$

$$E\left[\frac{(SSE)}{(q-1)(p-1)}\right] = E(MSSE) = \sigma_e^2$$

Hence under the null hypothesis

$$H_{0\alpha}: \alpha_1 = \alpha_2 = \ldots. = \alpha_p = 0;$$

$$H_{0\beta}: \beta_1 = \beta_2 = \ldots. = \beta_q = 0$$

$$E(MSST) = \sigma_e^2 \text{ and } E(MSSB) = \sigma_e^2$$

i.e. each of the mean sum of squares due to treatments and blocks gives an unbiased estimate of the error variance σ_e^2 under the null hypothesis $H_{0\alpha}$ and $H_{0\beta}$ respectively.

Example 1: There were 4 different makes of cars. A problem was posed to estimate the petrol consumption rates of the different makes of cars for suitable average speed and compare them. The following experiment could be conducted for an inference about the problem:

Five different cars of each four makes were chosen at random. The five cars of each make were put on road on 5 different days. The cars of A make run with different speeds on different days. The speeds were 25, 35, 50, 60 and 70 mph. Which car was to put on the road on which day and what speed it should have was determined through a chance mechanism subject to the above conditions of the experiment. The procedure was adopted for each of the makes of cars. For each car, the number of miles covered per gallon of petrol was observed. The observations are presented below:

Table: Miles per Gallon of Petrol

Makes of Car	**Speed of the cars in miles per hour (mph)**						**Average**
	25	35	50	60	70	Total	
A	20.6	19.5	18.1	17.9	16.0	92.1	18.42
B	19.5	19.0	15.6	16.7	14.1	84.9	16.98
C	20.5	18.5	16.3	15.2	13.7	84.2	16.84
D	16.2	16.5	15.7	14.8	12.7	75.9	15.18
Total	76.8	73.5	65.7	64.6	56.5	337.1	

Carry out the analysis of the given RBD.

Solution: Here the makes of the cars are the treatments and the other controlled factor is the speed, the variance for which has been eliminated through the design which is thus actually a randomised block design with the speeds as blocks. The specific cars used, the effects of the days, drivers and possibly some other effects contributed to the error variance.

Here,

Correction Factor (CF) = $\frac{(337.1)^2}{20} = 5681.82$

Raw Sum of Squares = $(20.6)^2 + (19.5)^2 + ... + (13.7)^2 + (12.7)^2 = 5781.41$

Total Sum of Squares (TSS) = 5681.41 - 5681.82 = 99.59

Sum of Squares due to Speed (SSS)

$$= \frac{(76.8)^2 + (73.5)^2 + ... + (64.6)^2 + (56.5)^2}{4} - CF$$

$$= 66.04$$

Sum of Squares due to Makes (SSM)

$$= \frac{(92.1)^2 + (84.9)^2 + (84.2)^2 + (75.9)^2}{5} - CF$$

$$= 28.78$$

Sum of Squares due to Errors (SSE)

$$= TSS - SSS - SSM$$

$$= 99.59 - 66.04 - 28.78 = 4.77$$

Analysis of Variance Table

Source of Variation	DF	SS	MSS	Variance Ratio	
				Calculated	Tabulated
Speeds	4	66.04	16.57	41.27	3.26
Treatments (Makes)	3	28.78	9.59	23.97	3.49
Error	12	4.77	0.40		
Total	19	99.59			

In both the cases either for speeds or for makes, calculated value of F is greater than tabulated value of F at 5% level of significance and thus null hypothesis is rejected.

In the above experiment, we are interested only on makes so multiple camparison test will be applied for different makes.

Mean number of miles per gallon for different Makes

	Makes		
$\bar{A}$	$\bar{B}$	$\bar{C}$	$\bar{D}$
18.42	16.98	16.84	15.18

$$SE = \sqrt{\frac{2MSSE}{5}} = \sqrt{\frac{2 \times 0.40}{5}} = 0.40$$

Critical difference at 1 % level of significance

$$CD = t_{\alpha/2} \text{ (for error df)} \times SE = 3.055 \times 0.40 = 1.22$$

The initial difference indicates that the Make A is significantly better than all the other Makes.

Pair of Treatments	Difference	CD	Inference
A, B	$\lvert\overline{A} - \overline{B}\rvert = 1.44$	1.22	Significant
A, C	$\lvert\overline{A} - \overline{C}\rvert = 1.58$	1.22	Significant
A, D	$\lvert\overline{A} - \overline{D}\rvert = 3.24$	1.22	Significant
B, C	$\lvert\overline{B} - \overline{C}\rvert = 0.14$	1.22	Insignificant
B, D	$\lvert\overline{B} - \overline{D}\rvert = 1.8$	1.22	Significant
C, D	$\lvert\overline{C} - \overline{D}\rvert = 1.66$	1.22	Significant

Example 2: Carryout the analysis of the following design:

Varieties	Blocks			
	I	**II**	**III**	**IV**
A	7	16	10	11
B	14	15	15	14
C	8	16	7	11

Solution: Let us find the block and variety totals by the following table:

Varieties	Blocks				Total
	I	**II**	**III**	**IV**	
A	7	16	10	11	44
B	14	15	15	14	58
C	8	16	7	11	102
Total	29	47	32	36	144

Correction Factor (CF) = $\frac{(144)^2}{12} = 1728$

Raw Sum of Squares (RSS) = $(7)^2 + (14)^2 + \ldots + (14)^2 + (11)^2 = 1858$

Total Sum of Squares (TSS) = $1858 - 1728 = 130$

Block Sum of Squares (SSB) $= \frac{(29)^2 + (47)^2 + (32)^2 + (36)^2}{3} - CF$

$= \frac{841 + 2209 + 1024 + 1296}{3} - 1728$

$= 1790 - 1728 = 62$

Variety Sum of Squares (SSV) $= \frac{(44)^2 + (58)^2 + (42)^2}{4} - CF$

$= \frac{1936 + 3364 + 1764}{4} - 1728$

$= 1766 - 1728 = 38$

Sum of Squares due to Error (SSE) $= TSS - SSV - SSB$

$= 130 - 62 - 38 = 30$

ANOVA Table

Source of Variation	DF	SS	MSS	Variance Ratio	
				Calculated	Tabulated
Variety	2	38	19	3.8	5.14
Blocks	3	62	20.67	4.13	4.76
Error	6	30	5		
Total	11	130			

In both these cases either for varieties or for blocks, calculated value of F is less than tabulated value of f at 5% level of significance and thus null hypothesis is accepted and inferred that variety effect and block effect are insignificant.

E1) Carryout the analysis of following design:

Blocks			
I	**II**	**III**	**IV**
A 8	**C** 10	**A** 6	**B** 10
C 12	**B** 8	**B** 9	**A** 8
B 10	**A** 8	**C** 10	**C** 9

10.4 MISSING PLOTS TECHNIQUE IN RBD

Sometimes observations from one or more experimental units are not found (missing) due to some unavoidable causes. There may be some unforeseen causes for example in agricultural experiments damage by animal or pets, in animal experiment any animal may die or observations from one or more plot

is excessively large as compared to other plots and thus accuracy of such observation is often in doubt. In such situations, these observations are omitted and treated as missing.

In case of missing observations, analysis is done by estimating the missing observation. This type of analysis was given by Yates (1937) and it is known as missing plot technique.

10.4.1 One Missing Plot

Suppose without loss of generality that observation for treatment 1 in block 1 i.e. y_{11} is missing and let it is Y, then the observations for a RBD may be represented as below:

	T_1	T_2	…..	T_i	…	T_p	Total
B_1	y_{11}=Y	y_{21}	…	y_{i1}	…	y_{p1}	B_1' + Y
B_2	y_{12}	y_{22}	…	y_{i2}	…	y_{p2}	B_2
…	…	…	…	…	…	…	…
B_j	y_{1j}	y_{2j}	…	y_{ij}	…	y_{pj}	B_j
…	…	…	…	…	…	…	…
B_q	y_{1q}	y_{2q}	…	y_{iq}	…	y_{pq}	B_q
Total	T_1' + Y	T_2	….	T_i	…	T_p	G' + Y

where,

$B_1^{'}$ = total of all available (p −1) observations in 1st block

$T_1^{'}$ = total of all available (q −1) observations in 1st treatment.

$G^{'}$ = total of all available (pq −1) observations

On the basis of these totals we calculate different SS's as follows:

Sum of Squares for Blocks (SSB) = $\dfrac{(B_1^{'}+Y)^2+\sum_{j=2}^{q}B_j^2}{p}-\dfrac{(G'+Y)^2}{pq}$

Sum of Squares for Treatments (SST) = $\dfrac{(T_1^{'}+Y)^2+\sum_{i=2}^{p}T_i^2}{q}-\dfrac{(G'+Y)^2}{pq}$

Total Sum of Squares (TSS) = $\sum_i\sum_j y_{ij}^2+Y^2-\dfrac{(G'+Y)^2}{pq}$ where $(i,j)\neq(1,1)$

Sum of Squares due to Error (SSE) = TSS – SSB – SST

$$SSE = Y^2+\frac{(G'+Y)^2}{pq}-\frac{(B_1^{'}+Y)^2}{p}-\frac{(T_1^{'}+Y)^2}{q}+\text{terms not involving Y}$$

For obtaining the value of Y, we minimize the sum of squares due to error with respect to Y. This is obtained by solving the equation

$$\frac{\partial(SSE)}{\partial Y} = 2Y + \frac{2(G'+Y)}{qp} - \frac{2(B_1' + Y)}{p} - \frac{2(T_1' + Y)}{q} = 0$$

$$\Rightarrow Y + \frac{Y}{pq} - \frac{Y}{p} - \frac{Y}{q} = \frac{T_1'}{q} + \frac{B_1'}{p} - \frac{G'}{pq}$$

$$\Rightarrow \frac{Y(pq + 1 - q - p)}{pq} = \frac{pT_1' + qB_1' - G'}{pq}$$

$$\hat{Y} = \frac{pT_1' + qB_1' - G'}{(p-1)(q-1)}$$

$\hat{Y}$ is the least square estimate of the yield of the missing plot. The value of Y is inserted in the original table of yield and ANOVA is performed in the usual way except that for each missing observation 1 df is subtracted from total and consequently from error df.

10.4.2 Two Missing Plots

For two missing values, we convert the problem into one missing value by putting any value say the overall mean or mean of the available values of that block for which one value is missing or mean of the available values of that replicate in any missing cell and obtain the estimate of the second missing value by the above prescribed estimation formula. Then we put the estimate of this second missing value and estimate the first missing value for which originally mean was taken. We go on repeating the same procedure until we obtain two successive estimates which are not materially different. Method is illustrated below with examples.

Example 3: In the following data two values are missing. Estimate these values by Yates method and analyse:

Treatments	**Blocks**		
	I	**II**	**III**
A	12	14	12
B	10	y	8
C	x	15	10

Solution: We convert the two missing plots problems into one missing plot problem, for which we take the average of the values of I block in which x is missing. This average is (10+12)/2 = 11. Thus, the estimate of x is taken to be x_1=11 and it is inserted in place of x and form the following table of totals:

Treatments	**Blocks**			**Total**
	I	**II**	**III**	
A	12	14	12	$T_A = 38$
B	10	y	8	$T_B = 18 + y$
C	11	15	10	Tc = 36
Total	$B_1 = 33$	$B_2 = 29 + y$	$B_3 = 30$	$G = 92 + y$

Thus, from the above table we get

$p = 3, q = 3, B_2' = 29, T_B' = 18, G' = 92$

Applying the missing estimation formula

$$\hat{Y} = \frac{pT_1' + qB_1' - G'}{(q-1)(p-1)} = \frac{3\times 18 + 3\times 29 - 92}{4}$$

$$= \frac{54+87-92}{4} = \frac{49}{4} = 12.25 \approx 12$$

Now the estimated value of y is taken to be $y_1 = 12$ and it is inserted in place of y and the following table of totals is formed by taking x unknown:

Treatments	**Blocks**			**Total**
	I	**II**	**III**	
A	12	14	12	$T_A = 38$
B	10	12	8	$T_B = 30$
C	x	15	10	$T_c = 25 + x$
Total	$B_1 = 22+x$	$B_2 = 41$	$B_3 = 30$	$G = 93 + x$

Thus from the above table we get $p = 3, q = 3, B_1' = 22, T_c' = 25, G' = 93$

Again applying the missing estimation formula

$$\hat{x} = \frac{3\times 25 + 3\times 22 - 93}{4}$$

$$= \frac{75+66-93}{4} = \frac{48}{4} = 12$$

Thus, $x_2 = 12$

Again using $x_2 = 12$, we estimate the second estimate of y i.e. y_2 for which

$B_2' = 29, T_B' = 18, G' = 92$

$$\hat{y} = \frac{3\times 18 + 3\times 29 - 93}{4}$$

$$= \frac{54+87-93}{4} = \frac{47}{4} = 11.75 \approx 12$$

We see that the second estimate of y i.e. y_2 is not materially different from y_1.

Thus, we take the estimated values of $\hat{x} = 12$ and $\hat{y} = 12$. Inserting both the estimated values of x and y we get the following observations:

Treatments	Blocks			Total
	I	II	III	
A	12	14	12	$T_A = 38$
B	10	12	8	$T_B = 30$
C	12	15	10	Tc = 37
Total	$B_1 = 34$	$B_2 = 41$	$B_3 = 30$	G = 105

Correction Factor (CF) = $\frac{(105)^2}{9} = \frac{11025}{9} = 1225$

Raw Sum of Square (RSS) = $(12)^2 + (10)^2 + ... + (8)^2 + (10)^2 = 1261$

Total Sum of Squares (TSS) = $1261 - 1225 = 36$

Treatment Sum of Squares (SST) = $\frac{(38)^2 + (30)^2 + (37)^2}{3} - CF$

$= \frac{1444 + 900 + 1369}{3} - 1225$

$= \frac{3713}{3} - 1225 = 1237.67 - 1225$

$= 12.67$

Block Sum of Squares (SSB) = $\frac{(34)^2 + (41)^2 + (30)^2}{3} - CF$

$= \frac{1156 + 1681 + 900}{3} - 1225$

$= 1245.67 - 1225 = 20.67$

Error Sum of Squares (SSE) = TSS − SST − SSB

$= 36 - 12.67 - 20.67 = 2.66$

ANOVA Table

Source of Variation	DF	SS	MSS	Variance Ratio	
				Calculated	Tabulated
Treatments	3 - 1=2	12.67	6.34	4.77	9.55
Blocks	3 - 1=2	20.67	10.34	7.77	9.55
Error	4 - 2=2	2.66	1.33		
Total	8 - 2 =6				

In case of both treatments and blocks, calculated value of F is less than tabulated value of F at 5% level of significance, thus treatment and block means are not significantly different.

E2) For the given data the yield of the treatment C in 2nd block is missing. Estimate the missing value and analyse the data:

Blocks	Treatments			
	A	**B**	**C**	**D**
I	105	114	108	109
II	112	113	Y	112
III	106	114	105	109

10.5 SUITABILITY OF RBD

1. The RBD is suitable in the situations where it is possible to divide the experimental material into a number of blocks. If it is not possible to divide the experimental material, RBD cannot be used.

2. The RBD is suitable only when the number of treatments is small because as the number of treatments increases, the block size also increases and it disturbs the homogeneity of the block.

3. RBD is suitable only when experimental material is heterogeneous with respect to one factor only. If there is two-way heterogeneity, LSD is used.

10.5.1 Advantages and Disadvantages of RBD

Advantages of RBD

The RBD has many advantages over other designs. Some of them are listed below:

1. It is a flexible design. It is applicable to moderate number of treatments. If extra replication is necessary for some treatment, this may be applied to more than one unit (but to the same number of units) per block.

2. Since all the three principles of design of experiments are used, the conclusions drawn from RBD are more valid and reliable.

3. If data from individual units be missing then, analysis can be done by estimating it.

4. This is the most popular design in view of its simplicity, flexibility and validity. No other design has been used so frequently as the RBD.

5. This design has been shown to be more efficient or accurate than CRD, for most types of experimental work. The elimination of block sum of squares from error sum of squares, usually results in a decrease of error sum of squares.

6. Analysis is simple and rapid.

Disadvantages of RBD

1. The main disadvantage of RBD is that if the blocks are not internally homogeneous, then a large error term will result. In field experiments, it is usually observed that as the number of treatments increases, the block size increases and so one has lesser control over error.
2. The number of replications for each treatment is same. If replication is not same, the only remedy is to adopt CRD.
3. It cannot control two sided variation of experimental material simultaneously. That is why, it is not recommended when experimental material contains considerable variability.

10.6 SUMMARY

In this unit, we have discussed:

1. The randomised block design;
2. The layout of RBD;
3. The statistical analysis of RBD;
4. The missing plot techniques in RBD; and
5. The advantages and disadvantages as well as the suitability of RBD.

10.7 SOLUTIONS/ ANSWERS

E1) The given design is solved by method of analysis of variance for two-way classified data. The computation results are given as follows:

Correction Factor (CF) = 972

Raw Sum of Squares (RSS) = 998

Total Sum of Squares (TSS) = 26

Block Sum of Squares (SSB) = 4.67

Treatment Sum of Squares (SST) = 15.5

Error Sum of Squares (SSE) = 5.83

ANOVA Table

Source of Variation	DF	SS	MSS	Variance Ratio	
				Calculated	Tabulated
Variety	2	15.5	7.7	7.94	5.14
Blocks	3	4.67	1.56	1.61	4.76
Error	6	5.83	0.97		
Total	11	26			

In case of variety, calculated value of F is greater than the tabulated value at F at 5% level of siginificance, so we reject the null hypothesis and conclude that the treatment effect is significant, while for blocks, it

is not significant. For pairwise testing, we have to find the standard error of difference of two treatment means:

$$SE = \sqrt{\frac{2MSSE}{q}} = \sqrt{\frac{2 \times 0.97}{3}} = 0.80$$

Critical difference (CD) = SE $\times t_{\alpha/2}$ at error df

$$= 0.80 \times 2.447 = 1.96$$

Treatment means are

$$\bar{A} = \frac{30}{4} = 7.5, \qquad \bar{B} = \frac{37}{4} = 9.25 \qquad \bar{C} = \frac{41}{4} = 10.25$$

Pair of Treatments	Difference	CD	Inference
A, B	$\lvert\bar{A} - \bar{B}\rvert = 1.76$	1.96	Insignificant
A, C	$\lvert\bar{A} - \bar{C}\rvert = 2.75$	1.96	Significant
B, C	$\lvert\bar{B} - \bar{C}\rvert = 1.00$	1.96	Insignificant

E2) We have p = 3, q = 4, $B_3' = 213$, $T_2' = 337$, $G' = 1207$ and the value of $\hat{y} = 109$

Therefore,

Correction Factor = 144321.33

Raw Sum of Squares = 144442.00

Total Sum of Squares = 120.67

Treatment Sum of Squares = 76.67

Block Sum of Squares = 20.67

Error Sum of Squares = 23.33

ANOVA Table

Source of Variation	DF	SS	MSS	Variance Ratio	
				Calculated	Tabulated
Treatments	3 - 1 = 2	20.67	10.33	2.21	5.79
Blocks	4 - 1 = 3	76.67	25.55	5.48	5.41
Error	6 - 1 = 5	23.33	4.66		
Total	11 - 1=10	120.67			

Block means are not but treatment means are significantly different at 5% level of significance.

In the above experiment, we are interested only treatments, so multiple camparison test will be applied for different treatments.

For pairwise testing, we have to find the standard error of difference of two treatment means:

$$SE = \sqrt{\frac{2MSSE}{p}} = \sqrt{\frac{2 \times 4.66}{3}} = 1.76$$

$$CD = SE \times t_{\alpha/2} \text{ at error df}$$
$$= 1.76 \times 2.447 = 4.31$$

Treatment means are

$$\overline{A} = \frac{323}{3} = 107.67,\ \overline{B} = \frac{341}{3} = 113.67,\ \overline{C} = \frac{322}{3} = 107.33, \overline{D} = \frac{330}{3} = 110$$

Pair of Treatments	**Difference of Treatment Means**	**CD**	**Inference**
A, B	$\lvert\overline{A} - \overline{B}\rvert = 6.0$	4.31	Significant
A, C	$\lvert\overline{A} - \overline{C}\rvert = 0.3$	4.31	Insignificant
A, D	$\lvert\overline{A} - \overline{D}\rvert = 2.3$	4.31	Insignificant
B, C	$\lvert\overline{B} - \overline{C}\rvert = 6.3$	4.31	Significant
B, D	$\lvert\overline{B} - \overline{D}\rvert = 3.7$	4.31	Insignificant
C, D	$\lvert\overline{C} - \overline{D}\rvert = 2.7$	4.31	Insignificant

UNIT 11 LATIN SQUARE DESIGN

Structure

11.1 Introduction
Objectives
11.2 Layout of Latin Square Design (LSD)
11.3 Statistical Analysis of LSD
11.4 Missing Plots Technique in LSD
One Missing Plot
11.5 Suitability of LSD
11.6 Summary
11.7 Solutions / Answers

11.1 INTRODUCTION

We know that RBD is used when experimental material is heterogeneous with respect to one factor and this factor of variation is eliminated by grouping the experimental material into a number of homogeneous groups called blocks. This grouping can be carried one step forward and we can group the units in two ways, each way corresponding to a source of variation among the units, and get the LSD. In agricultural experiments generally fertility gradient is not always known and in such situations LSD is used with advantage. Then LSD eliminates the initial variability among the units in two orthogonal directions.

The latin square design represents, in some sense, the simplest form of a row-column design. It is used for comparing m treatments in m rows and m columns, where rows and columns represent the two blocking factors. Latin squares and their combinatorial properties have been attributed to Euler (1782). They were proposed as experimental designs by Fisher (1925, 1926), although De Palluel (l788) already utilized the idea of a 4×4 latin square design for an agricultural experiment (see Street and Street, 1987, 1988).

Layout and statistical analysis of latin square design are explained in Sections 11.2 and 11.3. Missing plots technique in LSD for one missing plot is described in Section 11.4 whereas the suitability of LSD is explored in Section 11.5.

Objectives

After studying this unit, you would be able to

- explain the latin square design;
- describe the layout of LSD;
- explain the statistical analysis of LSD;
- find out the missing plot in LSD; and
- explain the advantages and disadvantages of LSD.

11.2 LAYOUT OF LATIN SQUARE DESIGN (LSD)

Mathematically speaking, the latin square of order m is an arrangement of m latin letters in a square of m rows and m columns such that every latin letter occurs once in each row and once in each column, or more generally, the arrangement of m symbols in a $m \times m$ array such that each symbol occurs exactly once in each row and column. In the context of experimental design, the latin letters are the treatments. Latin squares exist for every m. A reduced latin square (or latin square in standard form) is one in which the first row and the first column are arranged in alphabetical order, for example, for m = 3,

A B C
B C A
C A B

This is the only reduced latin square. The number of squares that can be generated from a reduced latin square by permutation of the rows, columns, and letters is (m!). These are not necessarily all different. If all rows but the first and all columns are permuted, we generate m! (m − l)! different squares. From the reduced latin square of order 3 we can thus generate $3! \times (3-1)! = 12$ squares.

In LSD two restrictions are imposed by forming blocks in two orthogonal directions, row-wise and column-wise. Further in LSD the number of treatments equals the number of replications of the treatment. Let there are m treatments and each is replicated m times then the total number of experimental units needed for the designs are $m \times m$. These m^2 units are arranged in m rows and m columns. Then m treatments are allotted to these m^2 units at random subject to the condition that each treatment occurs once and only once in each row and in each column.

Selected Latin Squares

3×3	**4×4**			
	1	2	3	4
ABC BCA CAB	ABCD BADC CDBA DCAB	ABCD BCDA CDAB DABC	ABCD BDAC CADB DCBA	ABCD BADC CDAB DCBA
5×5	**6×6**		**7×7**	
ABCDE BAECD CDAEB DEBAC ECDBA	ABCDEF BFDCAE CDEFBA DAFECB ECABFD FEBADC		ABCDEFG BCDEFGA CDEFGAB DEFGABC EFGABCD FGABCDE GABCDEF	

For randomization purpose two-way heterogeneity is eliminated by means of rows and columns and a latin square of order $m \times m$ is picked up from the table of Fisher and Yates. After picking the latin square its rows and columns are randomised by the help of random numbers and this randomised square is superimposed on the arranged square.

11.3 STATISTICAL ANALYSIS OF LSD

Let y_{ijk} (i, j, k = 1, 2, …, m) denote the response from unit (plot in the filed experimentation) in the i^{th} row, j^{th} column and receiving the k^{th} treatment. The triple (i, j, k) assumes only m^2 of the possible m^3 values of a LSD selected by the experiment. If S represents the set of m^2 values, then symbolically (i, j, k) belongs to S. If a single observation is made per experimental unit, then the linear additive model is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + e_{ijk}; \quad (i, j, k) \in S$$

where, μ is the general mean effect, α_i , β_j and τ_k are the constants effects due to the i^{th} row, j^{th} column and k^{th} treatment respectively and e_{ijk} is the error effect due to random component assumed to be normally distributed with mean zero and variance σ_e^2 i.e. e_{ijk} follows (i.i.d.) $N(0, \sigma_e^2)$.

If we write that

$G = y_{...}$ = Grand total of all the m^2 observations.

$R_i = y_{i..}$ = Total for m observations in the i^{th} row.

$C_j = y_{.j.}$ = Total of the m observations in the j^{th} column.

$T_k = y_{..k}$ = Total of the m observations in the k^{th} treatment.

Then heuristically, we get

$$\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2 = \sum_i \sum_j \sum_k [(\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{..k} - \bar{y}_{...})$$

$(i, j, k) \in S$ $\qquad + (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})\ ]^2$

$$= m\sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 + m\sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2 + m\sum_k (\bar{y}_{..k} - \bar{y}_{...})^2$$

$$+ \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2$$

The product terms vanish since the algebraic sum of deviations from mean is zero. Thus

$$TSS = SSR + SSC + SST + SSE$$

where TSS is the total sum of squares and SSR, SSC, SST and SSE are sum of squares due to rows, columns, treatments and due to error respectively given by

$$TSS = \sum_{i,j,k \in S} (y_{ijk} - \bar{y}_{...})^2;$$

$$SSR = m\sum_i (\bar{y}_{i...} - \bar{y}_{...})^2 = S_R^2 \text{ (say)}$$

$$SSC = m\sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2 = S_C^2$$

$$SST = m\sum_{k}(\bar{y}_{..k} - \bar{y}_{...})^2 = S_T^2$$

and $SSE = S_E^2 = TSS - SSR - SSC - SST$

Hence, the Total sum of squares is partitioned into three sum of squares, whose degree of freedom add to the degree of freedom of TSS.

ANOVA Table for LSD

Source of Variation	DF	SS	MSS	Variance Ratio(F)
Treatments	m– 1	S_T^2	$MSST = S_T^2/(m-1)$	$F_T = \frac{MSST}{MSSE}$
Columns	m– 1	S_C^2	$MSSC = S_C^2/(m-1)$	$F_C = \frac{MSSC}{MSSE}$
Rows	m– 1	S_R^2	$MSSR = S_R^2/(m-1)$	$F_R = \frac{MSSR}{MSSE}$
Error	(m– 1)(m– 2)	S_E^2	$MSSE = S_E^2/(m-1)(m-2)$	
Total	m^2-1			

Under the null hypothesis,

For row effects $H_{0\alpha}: \alpha_1 = \alpha_2 = \ldots = \alpha_m = 0$

For column effects $H_{0\beta}: \beta_1 = \beta_2 = \ldots = \beta_m = 0$ and

For treatment effects $H_{0\tau}: \tau_1 = \tau_2 = \ldots = \tau_m = 0$

against the alternative that all α's, β's and τ's are not equal, the test statistics F_T, Fc, F_R follow F distribution with [(m – 1), (m – 1) (m – 2)] df, under the above null hypothesis.

Thus, $F_\alpha = F_\alpha$ [(m– 1), (m– 1) (m– 2)] be the tabulated value of F distribution with [(m– 1), (m– 1) (m– 2)] df at the level of significance α. Thus, if $F_R > F_\alpha$ we reject the null hypothesis $H_{0\alpha}$, otherwise accept the null hypothesis. Similarly, we can test for $H_{0\beta}$ and $H_{0\tau}$.

Remark 1: Efficiency of LSD over RBD

There may be two cases to judge the relative efficiency of LSD over RBD:

1. Relative efficiency of LSD over RBD, when rows are taken as blocks is

$$= \frac{MSSC + (m-1)MSSE}{m \times MSSE}$$

2. Relative efficiency of LSD over RBD, when columns are taken as blocks is

$$= \frac{MSSR + (m-1)MSSE}{m \times MSSE}$$

Remark 2: Efficiency of LSD over CRD

Relative efficiency of LSD over CRD is given by

$$= \frac{MSSR + MSSC + (m-1)MSSE}{(m+1)MSSE}$$

Example 1: The example of petrol consumption by different makes of cars for illustrating randomised block designs has been converted to one with 5 makes of cars to illustrate latin square design. The effects of day and driver on consumption rate have been eliminated in addition to the effect of speed by suitable modification of the experimental situation. For this purpose, 5 drivers were chosen and each driver was used on one of 5 days. On that day, he drove 5 cars each of different make and each car with a different speed. The arrangement of the drivers, speeds and makes was as in the following table:

		Speeds in miles per hour				
		25	35	50	60	70
Drivers and Days	D_1	B(19.5)	E(21.7)	A(18.1)	D(14.8)	C(13.7)
	D_2	D(16.2)	B(19.0)	C(16.3)	A(17.9)	E(17.5)
	D_3	A(20.6)	D(16.5)	E(19.5)	C(15.2)	B(14.1)
	D_4	E(22.5)	C(18.5)	D(15.7)	B(16.7)	A(16.0)
	D_5	C(20.5)	A(19.5)	B(15.6)	E(18.7)	D(12.7)

Solution: Here, D_i (i = 1, 2, 3, 4, 5) denotes the i^{th} driver driving in the i^{th} day. A, B, C, D and E denote the 5 Makes of the cars. In the first cell of the table indicates that a car of Make B was driven by D_1 on this day with a speed of 25 miles per hour. The alphabets in the other cells have similar meaning. The number of miles covered by a gallon of petrol is shown in bracket in each cell.

The design adopted is actually a latin square design with the makes of cars as treatments and the drivers and speeds are the two controlled factors representing rows and columns. The observations of the miles per hour have been analysed below as appropriate for the design.

Correction Factor = 7638.76

Sum of Squares due to Speeds = 7719.49 –7638.76 = 80.73

Sum of Squares due to Drivers = 7640.12 – 7638.76 = 1.36

Sum of Squares due to Makes = 7704.18 – 7638.76 = 65.42

Total Sum of Squares = 7792.70 – 7638.76 = 153.94

Error Sum of Squares = 153.94 – 80.73 – 1.36 – 65.42 = 6.43

Analysis of Variance Table

Sources of Variation	DF	SS	MS	F Calculated	F Tabulated
Speeds	4	80.73	20.18	37.37**	3.26
Drivers	4	1.36	0.34	0.63	
Makes	4	65.42	16.35	30.28**	
Error	12	6.43	0.54		
Total	24	153.94			
** highly significant					

Mean numbers of miles per gallon for the different makes arranged in order

$\bar{E}$	$\bar{A}$	$\bar{B}$	$\bar{C}$	$\bar{D}$
19.98	18.42	16.98	16.84	15.18

$$SE = \sqrt{\frac{2 \times MSSE}{5}} = \sqrt{\frac{2 \times 0.54}{5}} = 0.33$$

CD at 1 per cent = $3.055 \times 0.33 = 1.42$

The initial difference indicates that the Make E is significantly better than all the other Makes. Make A was better than B, C and D. Finally D is the worst.

Efficiency of Latin square

$$E\ (Drivers) = \frac{4 \times 0.34 + 0.54 \times 16}{20 \times 0.54} = \frac{.34 + 0.54 \times 4}{5 \times 0.54} = 0.93$$

$$E\ (Speeds) = \frac{4 \times 20.18 + 0.54 \times 16}{20 \times 0.54} = \frac{20.18 + 0.54 \times 4}{5 \times 0.54} = 8.27$$

The efficiency figures show that elimination of speed variation increased precision considerably while elimination of driver variation did not reduce error variance.

E1) Carry out ANOVA for the following design:

A 5	B 7	C 7	D 8	E 9
B 7	C 9	D 8	E 8	A 5
C 6	D 5	E 9	A 8	B 9
D 5	E 6	A 8	B 5	C 7
E 8	A 9	B 5	C 7	D 6

11.4 MISSING PLOTS TECHNIQUE IN LSD

As we have discussed in Section 10.4 of Unit 10, sometimes observations from one or more experimental units are not found (missing) due to some unavoidable causes. There may be some unforeseen causes for example in agricultural experiments damage by animal or pets, in animal experiment any animal may die or observations from one or more plot is excessively large as compared to other plots and thus accuracy of such observation is often in doubt. In such situations, these observations are omitted and treated as missing.

In case of missing observations, analysis is done by estimating the missing observation. This type of analysis was given by Yates (1937) and it is known as missing plot technique. As similar as in the RBD, we are now going to discuss the same in LSD in the following sub-section:

11.4.1 One Missing Plot

Suppose without loss of generality that in $m \times m$ latin square design the observation occurring in the first row, first column and receiving first treatment is missing. Let us assume that $y_{111}=Y$

$R_1^{'}$ = Total of all available (m – 1) observations in 1st row.

$C_1^{'}$ = Total of all available (m – 1) observations in 1st column.

$T_1^{'}$ = Total of all available (m – 1) observations receiving 1st treatment.

$G^{'}$ = Total of all available $(m^2 - 1)$ observations.

On the basis of these totals we calculate different sum of squares as follows:

$$\text{Sum of Squares for Rows (SSR)} = \frac{(R_1^{'}+Y)^2+\sum_{i=2}^{m} R_i^2}{m} - \frac{(G'+Y)^2}{m^2}$$

$$\text{Sum of Squares for Columns (SSC)} = \frac{(C_1^{'}+Y)^2+\sum_{j=2}^{m} C_j^2}{m} - \frac{(G'+Y)^2}{m^2}$$

$$\text{Sum of Squares for Treatments (SST)} = \frac{(T_1^{'}+Y)^2+\sum_{k=2}^{m} T_k^2}{m} - \frac{(G'+Y)^2}{m^2}$$

$$\text{Total Sum of Squares (TSS)} = \sum_i \sum_j \sum_k y_{ijk}^2 + Y^2 - \frac{(G'+Y)^2}{m^2}$$
$$(i, j, k) \neq (1,1,1)$$

Sum of Squares due to Error (SSE) = TSS – SSR – SSC – SST

$$SSE = Y^2 + \frac{2(G'+Y)^2}{m^2} - \frac{(R_1' + Y)^2}{m}$$

$$- \frac{(C_1' + Y)^2}{m} - \frac{(T_1' + Y)^2}{m} + \text{Terms not involving Y}$$

For obtaining the value of Y, we minimize the sum of squares due to error with respect to Y. This is obtained by solving the equation

$$\frac{\partial(SSE)}{\partial Y} = 2Y + \frac{4(G'+Y)}{m^2} - \frac{2(R_1' + Y)}{m} - \frac{2(C_1' + Y)}{m} - \frac{2(T_1' + Y)}{m} = 0$$

$$\Rightarrow Y + \frac{2Y}{m^2} - \frac{Y}{m} - \frac{Y}{m} - \frac{Y}{m} = \frac{R_1'}{m} + \frac{C_1'}{m} + \frac{T_1'}{m} - \frac{2G'}{m^2}$$

$$\Rightarrow \frac{Y(m^2 + 2 - 3m)}{m^2} = \frac{m(R_1' + C_1' + T_1') - 2G'}{m^2}$$

$$\hat{Y} = \frac{m(R_1' + C_1' + T_1') - 2G'}{(m-1)(m-2)}$$

$\hat{Y}$ is the least square estimate of the yield of the missing plot. The value of Y is inserted in the original table of yield and ANOVA is performed in the usual way except that for each missing observation 1 df is subtracted from total and consequently from error df.

Example 2: In the following data, one value is missing. Estimate this value and analyse the given data.

Column / Row	I	II	III	IV	Row Totals (R_i)
I	A 12	C 19	B 10	D 8	49
II	C 18	B 12	D 6	A 7	43
III	B 22	D Y	A 5	C 21	48+Y
IV	D 12	A 7	C 27	B 17	63
Column Totals (C_j)	64	38+Y	48	53	203+Y

Solution: Here $m = 4, R_3' = 48,\ C_2' = 38, T_4' = 26,\ G' = 203$

Applying the missing estimation formula

$$\hat{Y} = \frac{m\left(R_3' + C_2' + T_4'\right) - 2G'}{(m-1)(m-2)}$$

$$= \frac{4(48+38+26) - 2\times 203}{(4-1)(4-2)} = 7$$

Inserting the estimated value of Y, we get the following observations:

Column / Row	I	II	III	IV	Row Totals (R_i)
I	A 12	C 19	B 10	D 8	49
II	C 18	B 12	D 6	A 7	43
III	B 22	D 7	A 5	C 21	55
IV	D 12	A 7	C 27	B 17	63
Column Totals (C_j)	64	45	48	53	210

Correction Factor (CF) = $\frac{(210)^2}{16} = \frac{44100}{16} = 2756.25$

Raw Sum of Squares (RSS) = $(12)^2 + (18)^2 + \ldots + (21)^2 + (17)^2 = 3432$

Total Sum of Squares (TSS) = 3432– 2756.25 = 675.75

Row Sum of Squares (SSR) = $\frac{(49)^2 + (43)^2 + (55)^2 + (63)^2}{4} - CF$

$$= \frac{2401+1849+3025+3969}{4} - 2756.25 = 54.75$$

Column Sum of Squares (SSC) = $\frac{(64)^2 + (45)^2 + (48)^2 + (53)^2}{4} - CF$

$$= \frac{4096+2025+2304+2809}{4} - 2756.25 = 52.25$$

Treatment Sum of Squares (TSS) = $\frac{(31)^2 + (61)^2 + (85)^2 + (33)^2}{4} - CF$

$$= \frac{961+3421+7225+1089}{4} - 2756.25 = 417.75$$

Error Sum of Squares (SSE) = TSS– SSR – SSC – SST

= 675.75 – 54.75 – 52.25 – 417.75 = 151

ANOVA Table

Source of Variation	DF	SS	MSS	Variance Ratio		Conclusion
				Calculated	Tabulated	
Rows	4– 1=3	54.75	18.25	0.60	5.41	Insignificant
Columns	4– 1=3	52.25	17.42	0.58	5.41	Insignificant
Treatments	4– 1=3	417.75	139.25	4.61	5.41	Insignificant
Error	6– 1=5	151	30.20			
Total	15– 1 =14					

E2) Estimate the missing value in the following LSD and then carry out the analysis of variance test.

Column / Row	I	II	III	IV
I	A 8	C 18	B 11	D 8
II	C 16	B 10	D 7	A Y
III	B 12	D 10	A 6	C 20
IV	D 10	A 9	C 28	B 16

11.5 SUITABILITY OF LSD

The latin square design is used when the experimental material is heterogeneous with respect to two factors and this two-way heterogeneity is eliminated by means of rows and columns. In fact LSD can be applied to all those cases where either the variation in the experimental material is not known or is known in two mutually perpendicular directions. Thus, LSD is successfully used in industry, animal husbandry, biological and social sciences, piggeries, marketing, medical and educational fields, where it is desired to eliminate the two factor heterogeneity simultaneously.

11.5.1 Advantages and Disadvantages of LSD

Advantages of LSD

1. Since total variation is divided into three parts namely rows, columns and treatments, the error variance is reduced considerably. It happens due to the fact that rows and columns being perpendicular to each other, eliminates the two-way heterogeneity up to a maximum extent.

2. LSD is an incomplete three way layout. Its advantage over the complete three way layout is that instead of m^3 units only m^2 units are needed. Thus, a 4×4 LSD results in saving 64 – 16 = 48 observations over a complete three way layout.
3. The analysis creates no problem even if a missing observation exists.

Disadvantages of LSD

1. The fundamental assumption that there is no interaction between different factors may not be true in general.
2. The main limitation of LSD is the equality of number of rows to that of columns and treatments. If the layout of experimental material is not of square design then LSD cannot be used.
3. RBD can be accommodated in any shape of field whereas for LSD field should perfectly be a square.
4. For smaller number of treatments, say less than 5, the degree of freedom for error is very small and thus the results are not reliable. Even in case of 2×2 LSD, degree of freedom for error becomes zero. In such situations, either the number of treatments should be increased or the latin square should be repeated.
5. On the other side, if the number of treatments increases the size of latin squares increases and this causes a disturbance in heterogeneity.
6. Analysis of LSD becomes very much complicated if complete row or complete column is missing. Analysis of RBD is quite easy in such situations.

11.6 SUMMARY

In this Unit, we have discussed:

1. The Latin Square design;
2. The layout of LSD;
3. The method of statistical analysis of LSD;
4. The missing plots technique in LSD; and
5. The advantages and disadvantages of LSD.

11.7 SOLUTIONS / ANSWERS

E1) The analysis of the given design is done by the method of analysis of variance. The computation results are given as follows:

Correction factor (CF)	=	1239.04
Raw Sum of Squares	=	1292
Total Sum of Squares	=	52.92
Column Sum of Squares	=	4.56
Row Sum of Squares	=	4.96

Treatment Sum of Squares = 7.76

Error Sum of Squares = 35.68

Analysis of Variance Table

Sources of Variation	DF	SS	MSS	F
Rows	4	4.96	1.24	0.42
Columns	4	4.56	1.14	0.38
Treatment	4	7.76	1.94	0.65
Error	12	35.68	2.97	
Total	24	52.96		

Tabulated value of F (4, 12) = 3.26

Since the calculated value of F is much less than the tabulated value of F at 5% level of significance, we conclude that there is no significant difference between treatment means.

E2) Let the missing value is Y then we have

Column / Row	I	II	III	IV	Row Totals (R_i)
I	A 8	C 18	B 11	D 8	45
II	C 16	B 10	D 7	A Y	33 + Y
III	B 12	D 10	A 6	C 20	48
IV	D 10	A 9	C 28	B 16	63
Column Totals (C_j)	46	47	52	44 + Y	189 + Y

Here, $m = 4,\ R_2' = 33,\ C_4' = 44,\ T_1' = 23,\ G' = 189$

Applying the missing estimation formula

$$\hat{Y} = \frac{m\left(R_3' + C_2' + T_4'\right) - 2G'}{(m-1)(m-2)}$$

$$= \frac{4(33+44+23) - 2\times 189}{(4-1)(4-2)} = 3.66 \sim 4$$

Inserting the estimated value of Y, we get the following observations:

Column / Row	I	II	III	IV	Row Totals (R_i)
I	A 8	C 18	B 11	D 8	45
II	C 16	B 10	D 7	A 4	37
III	B 12	D 10	A 6	C 20	48
IV	D 10	A 9	C 28	B 16	63
Column Totals (C_j)	46	47	52	48	193

Correction Factor (CF) = $\frac{(193)^2}{16} = 2328.06$

Raw Sum of Squares (RSS) = $(8)^2 + (16)^2 + ... + (20)^2 + (16)^2 = 2895$

Total Sum of Squares (TSS) = 2895 – 2328.06 = 566.94

Row Sum of Squares (SSR) = $\frac{(45)^2 + (37)^2 + (48)^2 + (63)^2}{4} - CF$

$= \frac{2025 + 1369 + 2304 + 3969}{4} - 2328.06 = 88.69$

Column Sum of Squares (SSC) = $\frac{(46)^2 + (47)^2 + (52)^2 + (48)^2}{4} - CF$

$= \frac{2116 + 2209 + 2704 + 2304}{4} - 2328.06 = 5.19$

Treatment Sum of Squares (SST) = $\frac{(27)^2 + (49)^2 + (82)^2 + (35)^2}{4} - CF$

$= \frac{729 + 2401 + 6724 + 1225}{4} - 2328.06 = 441.69$

Error Sum of Squares (SSE) = TSS− SSR – SSC – SST

= 566.94 – 88.69 – 5.19 – 441.69

= 31.37

ANOVA Table

Source of Variation	DF	SS	MSS	Variance Ratio		Conclusion
				Calculated	Tabulated	
Rows	4−1=3	88.69	29.56	4.71	5.41	Insignificant
Columns	4−1=3	5.19	1.73	0.28	5.41	Insignificant
Treatments	4−1=3	441.69	147.23	23.48	5.41	Significant
Error	6−1=5	31.37	6.27			
Total	15−1 =14					

Since for treatment effect calculated value of F is greater than the tabulated value of F at 5% level of significance, so we conclude that the treatment effect is significant. For pairwise testing, find the standard error of difference of two treatment means.

$$SE = \sqrt{\frac{2MSSE}{m}} = \sqrt{\frac{2\times 6.27}{4}} = 1.77$$

Critical difference (CD) = $SE \times t_{\alpha/2}$ at error df

$$= 1.77 \times 2.571 = 4.55$$

Treatment means

$$\overline{A} = \frac{27}{4} = 6.75,\ \overline{B} = \frac{49}{4} = 12.25,\ \overline{C} = \frac{82}{4} = 20.5\ \&\ \overline{D} = \frac{35}{4} = 8.75$$

Pair of Treatments	Difference	CD	Inference
A, B	$\lvert\overline{A} - \overline{B}\rvert = 05.50$	4.55	Significant
A, C	$\lvert\overline{A} - \overline{C}\rvert = 13.75$	4.55	Significant
A, D	$\lvert\overline{A} - \overline{D}\rvert = 02.00$	4.55	Insignificant
B, C	$\lvert\overline{B} - \overline{C}\rvert = 08.25$	4.55	Significant
B, D	$\lvert\overline{B} - \overline{D}\rvert = 03.50$	4.55	Insignificant
C, D	$\lvert\overline{C} - \overline{D}\rvert = 11.75$	4.55	Insignificant

UNIT 12 FACTORIAL EXPERIMENTS

Structure

12.1 Introduction
Objectives
12.2 Factorial Experiments
2^2 Factorial Experiments
12.3 Statistical Analysis of 2^2 Factorial Experiments
Step for Analysis
12.4 Statistical Analysis of 2^3 Factorial Experiments
12.5 Summary
12.6 Solutions / Answers

12.1 INTRODUCTION

Factorial experiments are the experiments that investigate the effects of two or more factors or input parameters on the output response of a process. Factorial experiment design or simply factorial design is a systematic method for formulating the steps needed to successfully implement a factorial experiment. Estimating the effects of various factors on the output of a process with minimum number of observations is crucial to being able to optimize the output of the process.

In a factorial experiment, the effects of varying the levels of the various factors affecting the process output are investigated. Each complete trial or replication of the experiment takes into account all the possible combinations of the varying levels of these factors. Effective factorial design ensures that least number of experimental runs are conducted to generate the maximum amount of information about how input variables affect the output of a process.

The basic definitions and different types of factorial experiments are described in Section 12.2. In Section 12.3, the statistical analysis of 2^2 factorial experiment is explained whereas the statistical analysis of 2^3 factorial experiment is given in Section 12.4.

Objectives

After studying this unit, you would be able to

- define the factorial experiments;
- describe the 2^2 factorial experiments;
- explain the statistical analysis of 2^2 factorial experiments;
- describe the 2^3 factorial experiments; and
- explain the statistical analysis of 2^3 factorial experiments.

12.2 FACTORIAL EXPERIMENTS

In factorial experiments, if the effet of two factors with levels r and s on the output of a process are investigated, then one would need to run $r \times s$ treatment combinations to complete the experiment. For instance if the effect of two factors A (Fertilizer) and B (Irrigation) on the output of a process are investigated, and A has 3 levels (0 kg N/ha, 20 kg N/ha and 40 kg N/ha) while B has two levels (5 cm and 10 cm), then one would need to run 6 treatment combinations to complete the experiment, observing the process output for each of the following combinations:

0 kg N/ha-5 cm I,

0 kg N/ha-10 cm I,

20 kg N/ha-5 cm I,

20 kg N/ha-10 cm I,

40 kg N/ha-5 cm I,

40 kg N/ha-10 cm I,

The amount of change produced in the process output for the change in the level of the given factor is referred as the main effect of that factor. Table 1 shows an example of simple factorial experiment involving two factors with two levels each. The two levels of each factor may be denoted by 'low' and 'high' which are usually symbolized by '–' and '+' in factorial designs, respectively.

Table 1: A Simple 2^2 Factorial Experiment

	A(–)	A(+)
B(–)	20	40
B(+)	30	52

The main effect of a factor is basically the average change in the output response as that factor goes from '–'to '+'. Mathematically this is the average of two numbers: 1) the change in the output when the factor goes from low to high level as the other factor stays low, and 2) the change in the output when the factor goes from low to high level as the other factor stays high.

In the example given in Table 1, the output of the process is just 20 (lowest output) when both A and B are at their '–'level. While the output is maximum at 52 when both A and B are at their '+' level. The main effect of A is the average of the change in the output response when B stays at '–'as A goes from '–'to '+' or (40−20) = 20 and the change in the output response when B stays '+' as A goes from '–' to '+' or (52– 30) = 22. The main effect of A, therefore, is equal to $\frac{20+22}{2} = 21$.

Similarly, the main effect of B is the average change in output as it goes from '–' to '+', i.e. the average of 10 and 12, 11. Thus, the main effect of B in this process is 11. Here one can see that the factor A exerts a greater influence on

the output of the process, having a main effect of 21 versus factor B's main effect of only 11.

It must be noted that aside from main effects factors can likewise result in interaction effects. Interaction effects are changes in the process output caused by two or more factors that are interacting with each other. Large interactive effects can make the main effects insignificant, such that it becomes more important to pay attention to the interaction of the involved factors than to investigate them individually. In above table as effects of A (B) is not same at all the levels of B (A) hence, A and B are interacting, thus interaction is the failure of the differences in response to changes in levels of one factor, to retain the same order and magnitude of performance throughout all the levels of other factors or the factors are said to interact if the effect of one factor changes as the levels of other factor(s) changes.

If interaction exists which is fairly common, we should plan our experiment in such a way that they can be estimated and tested. It is clear that we cannot do this if we vary only one factor at a time. For this purpose, we must use multilevel, multifactor experiments.

The running of factorial combinations and the mathematical interpretation of the output responses of the process to such combinations is the essence of factorial experiments. It allows us to understand which factors affect the process most so that improvements may be graded towards these.

We may define factorial experiments as experiments in which the effects (main effects and interactions) of more than one factor are studied together. In general if there are n factors say F_1, F_2, …, F_n and i^{th} factor has s_i levels, i=1, 2, …, n, then total number of treatment combinations is $\prod s_i$. Factorial experiments are of two types.

Factorial experiments in which the number of levels of all the factors are same i.e. all s_i are equal, are called symmetrical factorial experiments and the experiments in which at least two of the s_i's are different are called as asymmetrical factorial experiments.

If there are p different varieties, then we shall say that there are p levels of the factor variety. Similarly, the second factor manure may have q levels i.e. there may be q different manures or different doses of the same manure. Then this factorial experiment is called $p \times q$ experiment and this example of variety (at p-levels) and manure (at q-levels) is related to asymmetrical factorial experiment.

12.2.1 2^2 Factorial Experiments

The simplest of the symmetrical factorial experiments are the experiments with each of two factors at 2 levels. If there are n factors each at 2 levels it is called as a 2^n factorial experiments, where power stands for number of factors and the base the level of each factor. Simplest of the symmetrical factorial experiments is the 2^2 factorial experiment i.e. 2 factors say A and B each at two levels 0 (low) and 1(high). There will be 4 treatment combinations which can be written as

$00 = a_0b_0 = 1$: A and B both at first (low) levels.

$10 = a_1b_0 = a$: A at second (high) level and B at first (low) level.

01 = a_0b_1 = b: A at first (low) level and B at second (high) level.

11 = a_1b_1 = ab: A and B both at second (high) level.

In a 2^2 experiment wherein r replicates were run for each treatment combination, the main and interaction effects of A and B on the output may be mathematically expressed as follows:

A = [ab + a – b – (1)] / 2r; (main effect of factor A)

B = [ab + b – a – (1)] /2r; (main effect of factor B)

AB = [ab+ (1) – a– b]/ 2r; (interaction effect of factor A and B)

12.3 STATISTICAL ANALYSIS OF 2^2 FACTORIAL EXPERIMENTS

If the two factors remain independent, then [ab+ (1) – a– b]/ 2r will be of the order of zero. If not then this will give an estimate of interdependence of the two factors and it is called the interaction between A and B. It is easy to verify that the interaction of the factor B with the factor A is BA which will be same as the interaction AB and hence the interaction does not depend on the order of the factors. It is also easy to verify that main effect of factor B, a contrast of the treatments totals is orthogonal to each of A and AB.

2^2 Factorial Experiment Pattern

RUN	Comb.	M	A	B	AB
1	(1)	+	–	–	+
2	a	+	+	–	–
3	b	+	–	+	–
4 = 2^2	ab	+	+	+	+

The signs for the main effects can be written according to the rule “Give a plus sign to each of the treatment means when the corresponding factor is at the high level and a minus sign where it is of the low level. Or give a plus sign to the treatment combinations containing the corresponding small letter and give a minus sign where the corresponding small letter is absent. For a two factor interaction, the signs are obtained by combining the corresponding signs of two main effects i.e. two opposite signs will give a minus sign and two identical signs will give a plus sign to the interaction.”

12.3.1 Steps for Analysis

1. The sum of squares (SS) due to treatments, replications (in case of RBD is used) due to rows and columns (in case a row-column design has been used), total sum of squares and error sum of squares is obtained as per established procedures. No replication sum of squares is required in case of a CRD. The treatment sum of squares is divided into different components i.e. main effects and interactions each with single df. The sum of squares due to these factorial effects is obtained by factorial effect

totals rather than treatment means. Factorial effect totals are may be defined as

$[A] = [ab] + [a] - [b] - [1]$

$[B] = [ab] - [a] + [b] - [1]$

$[AB] = [ab] - [a] - [b] + [1]$

where, each [ab], [a], [b], [1] is the total of outputs of each of the replicates of the treatment combinations ab, a, b,1 respectively.

2. For a 2^2 factorial experiment, the sum of squares due to a main effect or the interaction effect is obtained by dividing the square of the effect total by 4r. Thus

   Sum of squares due to main effect of A = $[A]^2/ 4r$, with 1 df

   Sum of squares due to main effect of B = $[B]^2/ 4r$, with 1 df

   Sum of squares due to interacton effect AB = $[AB]^2/ 4r$, with 1 df

   Mean squares (MS) is obtained by dividing each SS by corresponding

   degrees of freedom.

3. After obtaining the different SS's the usual ANOVA table is prepared and the different effects are tested against error mean square and conclusions drawn.

4. If we consider n factors then standard error (SE's) for main effects and two factors interaction:

   SE of difference between main effect means = $\sqrt{\frac{2MSSE}{r.2^{n-1}}}$,

Therefore, for n = 2, SE of difference between main effect means

$$= \sqrt{\frac{2MSSE}{r.2^{2-1}}}$$

SE of difference between A means at same level of B

= SE of difference between B means at same level of A

$$= \sqrt{\frac{2MSSE}{r.2^{n-2}}},$$

Therefore, for n = 2, SE of difference between A means at same level of B

$$= \sqrt{\frac{2MSSE}{r.2^{2-2}}}$$

In general, SE for testing the difference between means in case of r-factor interaction

$$= \sqrt{\frac{2MSSE}{r.2^{n-r}}}$$

The critical differences are obtained by multiplying the SE by the student's t value at α level of significance at error degree of freedom.

The ANOVA for 2^2 factorial experiments with r replications conducted using RBD is as follows:

ANOVA Table

Source of Variation	DF	SS	MSS	F
Between Replications	r – 1	SSR	MSSR= SSR/(r – 1)	MSSR/MSSE
Between Treatments	$2^2 – 1 = 3$	SST	MSST= SST/3	MSST/MSSE
A	1	SSA=$[A]^2$ /4r	MSSA = SSA/1	MSSA/MSSE
B	1	SSB=$[B]^2$ /4r	MSSB = SSB/1	MSSB/MSSE
AB	1	SSAB=$[AB]^2$ /4r	MSSAB = SSAB/1	MSSAB/MSSE
Error	3(r – 1)	SSE	MSSE = SSE / 3(r– 1)	
Total	r. $2^2 – 1$ = 4r– 1	TSS		

12.4 STATISTICAL ANALYSIS OF 2^3 FACTORIAL EXPERIMANT

Consider the case of 3 factors A, B and C each at two levels (0 and 1) i.e. 2^3 factorial experiment. There will be 8 treatment combinations which are written as

000 = $a_0b_0c_0$ = (1); A, B and C (all three) at first level.

100 = $a_1b_0c_0$ = a; A at second level and B and C are at first level.

010 = $a_0b_1c_0$ = b; A and C are both at first level and B at second level.

001 = $a_0b_0c_1$ = c; A and B both at first level and C at second level.

110 = $a_1b_1c_0$ = ab; A and B both at second level and C at first level.

101 = $a_1b_0c_1$ = ac; A and C both at second level and B at first level.

011 = $a_0b_1c_1$ = bc; A at first level and B and C at second level.

111 = $a_1b_1c_1$ = abc; A, B and C (all the three) at second level.

In a three factor experiment there are three main effects A,B and C; 3 first order or two factor interaction AB, AC, BC and one second order or three factor interaction ABC.

2^3 Factorial Experiment Pattern

RUN	Comb.	M	A	B	AB	C	AC	BC	ABC
1	(1)	+	–	–	+	–	+	+	–
2	a	+	+	–	–	–	–	+	+
3	b	+	–	+	–	–	+	–	+
4	ab	+	+	+	+	–	–	–	–
5	c	+	–	–	+	+	–	–	+
6	ac	+	+	–	–	+	+	–	–
7	bc	+	–	+	–	+	–	+	–
$8 = 2^3$	abc	+	+	+	+	+	+	+	+

$$\text{Main effect A} = \frac{1}{4}[(abc)-(bc)+(ac)-(c)+(ab)-(b)+(a)-(1)]$$

$$= \frac{1}{4}(a-1)(b+1)(c+1)$$

$$\text{AB} = \frac{1}{4}[(abc)-(bc)-(ac)+(c)+(ab)-(b)-(a)+(1)]$$

$$\text{ABC} = \frac{1}{4}[(abc)-(bc)-(ac)+(c)-(ab)+(b)+(a)-(1)]$$

Or equivalently,

$$\text{AB} = \frac{1}{4}(a-1)(b-1)(c+1)$$

$$\text{ABC} = \frac{1}{4}(a-1)(b-1)(c-1)$$

The method of representing the main effect or interaction as above is due to Yates and is very useful and quite straightforward. For example if the design is 2^4 then

$$\text{A} = (1/2^{3)}\,[(a-1)\,(b+1)\,(c+1)\,(d+1)]$$

$$\text{ABC} = (1/2^{3)}\,[(a-1)\,(b-1)\,(c-1)\,(d+1)]$$

In case of 2^n factorial experiment there will be $2^n = \nu$ treatment combinations with n main effects, '$\binom{n}{2}$' first order or two factor interactions, '$\binom{n}{3}$' second order or three factor interactions, '$\binom{n}{4}$' third order or four factor interactions and so on, '$\binom{n}{r}$', $(r-1)^{th}$ order or r factor interactions and '$\binom{n}{n}$', $(n-1)^{th}$ order or n factor interaction. Using these ν treatments combinations, the experiment may be laid out using any of the suitable experimental designs for example CRD or RBD, etc.

Steps for the analysis are same as explained for the 2^2 factorial experiments. ANOVA for 2^3 factorial experiment conducted in RBD with r replications is given by

ANOVA Table

Source of Variation	DF	SS	MSS	Variance Ratio (F)
Between Replications	(r – 1)	SSR	MSSR = SSR/ (r – 1)	MSSR/MSSE
Between Treatments	$2^3 – 1 = 7$	SST	MSST =SST/7	MSST/MSSE
A	1	SSA	MSSA = SSA/1	MSSA/MSSE
B	1	SSB	MSSB = SSB/1	MSSB/MSSE
AB	1	SSAB	MSSAB=SSAB/1	MSSAB/MSSE
C	1	SSC	MSSC=SSC/1	MSSC/MSSE
AC	1	SSAC	MSSAC=SSAC/1	MSSAC/MSSE
BC	1	SSBC	MSSBC=SSBC/1	MSSBC/MSSE
ABC	1	SSABC	MSSABC=SSABC/1	MSSABC/MSSE
Error	7(r – 1)	SSE	MSSE=SSE/7(r −1)	
Total	r. $2^3 −1 = 8r – 1$	TSS		

Similarly, ANOVA table for 2^n factorial experiment can be made.

Example 1: Analyse the data of a 2^3 factorial experiment conducted using a RBD with three replications. The three factors were fertilizers viz. Nitrogen (N), Phosphorus (P) and Potassium (K). The purpose of the experiment is to determine the effect of different kinds of fertilizers on potato crop yield. The yields under 8 treatment combinations for each of the three randomised blocks are given below:

Block I

npk	(1)	k	np	p	n	nk	pk
450	101	265	373	312	106	291	391

Block II

p	nk	k	np	(1)	npk	pk	n
324	306	272	338	106	449	407	89

Block III

p	npk	nk	(1)	n	k	pk	np
323	471	334	87	128	279	423	324

Solution: The data arranged in following table

Blocks ↓	**Treatment Combinations**								**Total**
	(1)	**n**	**p**	**np**	**k**	**nk**	**pk**	**npk**	
B_1	101	106	312	373	265	291	391	450	2289
B_2	106	89	324	338	272	306	407	449	2291
B_3	87	128	323	324	279	334	423	471	2369
Total	294	323	959	1035	816	931	1221	1370	6949
	(T_1)	(T_2)	(T_3)	(T_4)	(T_5)	(T_6)	(T_7)	(T_8)	(G)

Grand Total = 6949

Number of observations (N) = $r \times 2^n = 3 \times 2^3 = 3 \times 8 = 24$

Correction Factor (CF) = $\frac{G^2}{N} = \frac{(6949)^2}{24} = 2012025.042$

Total Sum of Squares (TSS) = Raw Sum of Squares – CF

$$= (101^2 + 106^2 + ... + 449^2 + 471^2) - CF$$

$$= 352843.958$$

Block (Replication) Sum of Squares (SSB) = $\sum_{j=1}^{r} \frac{B_j^2}{2^3} - CF$

$$= \frac{[(2289)^2 + (2291)^2 + (2369)^2]}{8} - CF$$

$$= 520.333$$

Treatment Sum of Squares (SST) = $\sum_{i=1}^{v} \frac{T_i^2}{r} - CF$

$$= \frac{[(294)^2 + (323)^2 + (959)^2 + (1035)^2 + (816)^2 + (931)^2 + (1221)^2 + (1370)^2]}{3} - CF$$

$$= \frac{7082029}{3} - 2012025.042$$

$$= 348651.291$$

Error Sum of Squares (SSE) = TSS – SSB – SST

= 352843.958 - 520.333 -348651.291

= 3672.334

Main effects totals and interactions totals are obtained as follows:

[N] = [npk] - [pk] + [nk] – [k] + [np] – [p] + [n] – [1] = 369

[P] = [npk] + [pk] – [nk] – [k] + [np] + [p] – [n] – [1] = 2221

[K] = [npk] + [pk] + [nk] + [k] - [np] – [p] – [n] – [1] =1727

[NP] = [npk] – [pk] – [nk] + [k] + [np] - [p] – [n] + [1] = 81

[NK] = [npk] – [pk] + [nk] – [k] – [np] + [p] – [n] + [1] = 159

[PK] = [npk] + [pk] – [nk] – [k] - [np] + [p] + [n] + [1] = -533

[NPK] = [npk] – [pk] – [nk] + [k] - [np] + [p] + [n] – [1] = -13

$$\text{Factorial effects} = \frac{\text{Fcatorial effect total}}{\text{r.}2^{n-1}(=12)}$$

$$\text{Factorial effect Sum of Squares} = \frac{(\text{Fcatorial effect total})^2}{\text{r.}2^{n}(=24)}$$

Here factorial effects,

N = 30.750, P = 185.083, K = 143.917, NP = 6.750, NK = 13.250,

PK = − 44.417, NPK = − 1.083

Sum of Squares due to N = 5673.375

Sum of Squares due to P = 202235.042

Sum of Squares due to K = 124272.042

Sum of Squares due to NP = 273.375

Sum of Squares due to NK = 1053.375

Sum of Squares due to PK = 11837.042

Sum of Squares due to NPK = 7.042

Mean sum of squares is obtained by dividing the sum of squares by their respective df.

ANOVA Table

Source of Variation	DF	SS	MSS	Variance Ratio (F)
Between Replications	(r −1) = 2	520.333	260.167	0.9918
Between Treatments	2^3 −1 = 7	348651.291	49807.327	189.8797
N	(2 −1) = 1	5673.375	5673.375	21.6285
P	1	205535.042	205535.042	783.5582
K	1	124272.042	124272.042	473.7606
NP	1	273.375	273.375	1.0422
NK	1	1053.375	1053.375	4.0158
PK	1	11837.041	11837.041	45.1262
NPK	1	7.041	7.041	0.0268
Error	7(r−1)= 14	3672.337	262.310	
Total	r. 2^3−1 = 23	352843.958		

Standard Error of difference between main effect means

$$= \sqrt{\frac{\text{MSSE}}{\text{r.}2^{n-2}}} = 8.098$$

SE of difference between N means at same level of P or K

= SE of difference between P (or K) means at same level of N

= SE of difference between P means at same level of K

= SE of difference between K means at same level of P

$$= \sqrt{\frac{\text{MSSE}}{\text{r.}2^{n-3}}} = 11.4523,$$

And $t_{0.05}$ at 14 df = 2.145. Accordingly Critical difference (CD) can be calculated.

E1) An experiment was planned to study the effect of Sulphate, Potash and Super Phosphate on the yield of potatoes. All the combinations of 2 levels of Super Phosphate [0 cent (p_0) and 5 cent (p_1)/ acre] and two levels of Sulphate and Potash [0 cent (k_0) and 5 cent (k_1)/acre] were studied in a randomised block design with 4 replications each. The (1/70) yields [lb. per plot = (1/70) acre] obtained are given in table below:

Blocks	Yields (lbs per plot)			
I	(1) 23	k 25	p 22	kp 38
II	p 40	(1) 26	k 36	kp 38
III	(1) 29	k 20	pk 30	p 20
IV	kp 34	k 31	p 24	(1) 28

Analyse the data and give your conclusions.

12.4.1 Advantages and Disadvantages of Factorial Experiments

Advantages

1. To investigate the interactions of factors. Single factor experiments provide a disorderly and incomplete picture.
2. In exploratory work for quick determination of which factors are independent and can therefore be more fully analyzed in separate experiments.
3. To lead to recommendations that must extend over a wide range of conditions.

Disadvantages

1. Large numbers of combination are required to study several factors at several levels and need a large sized experiment: 7 factors at 3 levels requires 2187 combinations.
2. Large numbers of factor, complicate the interpretation of high order interactions.

12.5 SUMMARY

In this unit, we have discussed:

1. The factorial experiments;
2. The layout of 2^2 factorial experiments;
3. The statistical analysis of 2^2 factorial experiments;
4. The layout of 2^3 factorial experiments, and
5. The statistical analysis of 2^3 factorial experiments.

12.6 SOLUTIONS/ANSWERS

E1) Correction Factor (CF) = 0

Raw Sum of Squares = 660

Total Sum of Squares = 660

Block Sum of Squares = 232.5

Treatment Sum of Squares = 198

Error Sum of Squares = 229.5

Sum of Squares due to k = 100

Sum of Squares due to p = 49

Sum of Squares due to kp = 49

ANOVA Table

Source of Variation	DF	SS	MSS	Variance Ratio	
				Calculated	Tabulated
Treatments	3	198	77.5	3.04	3.86
Blocks	3	232.5	66	2.59	3.86
k	1	100	100	3.92	5.12
p	1	49	49	1.92	5.12
kp	1	49	49	1.92	5.12
Error	6	229.5	25.5		
Total	11	660			

Calculated value of F is less than the corresponding tabulated value, so there are no significant main or interaction effects present in the experiment. The blocks as well as treatments do not differ considerably.

TABLE: The F Table

Value of F Corresponding to 5% (Normal Type) and 1% (Bold Type) of the Area in the Upper Tail

Degrees of Freedom: (Denominator)	Degrees of Freedom (Numerator) 1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	254
	4,052	**4,999**	**5,403**	**5,625**	**5,764**	**5,859**	**5,928**	**5,981**	**6,022**	**6,056**	**6,082**	**6,106**	**6,142**	**6,169**	**6,208**	**6,234**	**6,258**	**6,366**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.50
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.34**	**99.36**	**99.38**	**99.40**	**99.41**	**99.42**	**99.43**	**99.44**	**99.45**	**99.46**	**99.47**	**99.50**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.5
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**	**26.92**	**26.83**	**26.69**	**26.60**	**26.50**	**26.12**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.63
	22.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**	**14.24**	**14.15**	**14.02**	**13.93**	**13.83**	**13.46**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.40	4.36
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.27**	**10.15**	**10.05**	**9.96**	**9.89**	**9.77**	**9.68**	**9.55**	**9.47**	**9.38**	**9.02**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.67
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**	**7.60**	**7.52**	**7.39**	**7.31**	**7.23**	**6.88**
7	5.59	4.47	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.23
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**	**6.35**	**6.27**	**6.15**	**6.07**	**5.98**	**5.65**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	2.93
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**	**5.56**	**5.48**	**5.36**	**5.28**	**5.20**	**4.86**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.71
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**	**5.00**	**4.92**	**4.80**	**4.73**	**4.64**	**4.31**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.54
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**	**4.60**	**4.52**	**4.41**	**4.33**	**4.25**	**3.91**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.40
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**	**4.29**	**4.21**	**4.10**	**4.02**	**3.94**	**3.60**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.30
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**	**4.05**	**3.98**	**3.86**	**3.78**	**3.70**	**3.36**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.21
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**	**3.85**	**3.78**	**3.67**	**3.59**	**3.51**	**3.16**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.13
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**	**3.43**	**3.34**	**3.00**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.07
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**	**3.29**	**3.20**	**2.87**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.01
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**	**3.45**	**3.37**	**3.25**	**3.18**	**3.10**	**2.75**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	1.96
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.95**	**3.52**	**3.45**	**3.35**	**3.27**	**3.16**	**3.08**	**3.00**	**2.65**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	1.92
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**	**3.27**	**3.19**	**3.07**	**3.00**	**2.91**	**2.57**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	1.88
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**	**3.19**	**3.12**	**3.00**	**2.92**	**2.84**	**2.49**
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.84
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**	**3.13**	**3.05**	**2.94**	**2.86**	**2.77**	**2.42**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.81
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**	**3.07**	**2.99**	**2.88**	**2.80**	**2.72**	**2.36**

TABLE (Continued)

Degrees of Freedom: Denominator	Degrees of Freedom: Numerator																	
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	∞
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.23	2.23	2.18	2.13	2.07	2.03	1.98	1.78
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**	**3.02**	**2.94**	**2.83**	**2.75**	**2.67**	**2.31**
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	3.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.76
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**	**2.97**	**2.89**	**2.78**	**2.70**	**2.62**	**2.26**
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.73
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**	**2.93**	**2.85**	**2.74**	**2.66**	**2.58**	**2.21**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.71
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**	**2.89**	**2.81**	**2.70**	**2.62**	**2.54**	**2.17**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.69
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**	**2.86**	**2.77**	**2.66**	**2.58**	**2.50**	**2.13**
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.67
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**	**2.83**	**2.74**	**2.63**	**2.55**	**2.47**	**2.10**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.65
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**	**2.80**	**2.71**	**2.60**	**2.52**	**2.44**	**2.06**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.64
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**	**2.77**	**2.68**	**2.57**	**2.49**	**2.41**	**2.03**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.62
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**	**2.74**	**2.66**	**2.55**	**2.47**	**2.38**	**2.01**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.00
	6.64	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**	**2.07**	**1.99**	**1.87**	**1.79**	**1.69**	**1.00**

Indira Gandhi National Open University
School of Sciences

MST - 005
Statistical Techniques

Block

4

RANDOM NUMBER GENERATION AND SIMULATION TECHNIQUES

UNIT 13
Random Number Generation for Discrete Variables **5**

UNIT 14
Random Number Generation for Continuous Variables **21**

UNIT 15
Simulation Techniques **35**

UNIT 16
Applications of Simulation **43**

Curriculum and Course Design Committee

Prof. K. R. Srivathsan
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Parvin Sinclair
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Geeta Kaicker
Director, School of Sciences
IGNOU, New Delhi

Prof. R. M. Pandey
Department of Bio-Statistics
All India Institute of Medical Sciences
New Delhi

Prof. Jagdish Prasad
Department of Statistics
University of Rajasthan, Jaipur

Prof. Rahul Roy
Maths and Stat. Unit
Indian Statistical Institute, New Delhi

Dr. Diwakar Shukla
Department of Mathematics and Statistics
Dr. Hari Singh Gaur University, Sagar

Prof. G. N. Singh
Department of Applied Mathematics
I S M University, Dhanbad

Prof. Rakesh Srivastava
Department of Statistics
M. S. University of Baroda, Vadodara

Dr. Gulshan Lal Taneja
Department of Mathematics
M. D. University, Rohtak

Faculty Members, School of Sciences, IGNOU

Statistics
Dr. Neha Garg
Dr. Nitin Gupta
Mr. Rajesh Kaliraman
Dr. Manish Trivedi

Mathematics
Dr. Deepika Garg
Prof. Poornima Mital
Prof. Sujatha Varma
Dr. S. Venkataraman

Block Preparation Team

Content Editor
Prof. Rakesh Srivastava
Department of Statistics
M. S. University of Baroda, Vadodara

Course Writer
Prof. G. K. Shukla
Decision Science Group
Indian Institute of Management, Lucknow

Language Editor
Dr. Parmod Kumar
School of Humanities, IGNOU

Formated By
Dr. Manish Trivedi
Mr. Prabhat Kumar Sangal
School of Sciences, IGNOU

Secretarial Support
Mr. Deepak Singh

Programme and Course Coordinator: Dr. Manish Trivedi

Production

Mr. Y. N. Sharma, SO (P.)
School of Sciences, IGNOU

Acknowledgement: We gratefully acknowledge Prof. Geeta Kaicker, Director, School of Sciences for her great support and guidance.

December, 2011

ISBN – 978-81-266-5787-2

Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by Director, School of Sciences.

Laser Type set by: Rajshree Computers, V-166A, Bhagwati Vihar, (Near Sector-2, Dwarka), Uttam Nagar, New Delhi-110059.

Printed at: Gita Offset Printers Pvt. Ltd., C-90, Okhla Industrial Area, Phase-I, New Delhi-110020.

RANDOM NUMBER GENERATION AND SIMULATION TECHNIQUES

Simulation is becoming a very important tool for analysis of a wide variety of problems whose exact solution is very difficult, expensive or time taking. In the last few years it has been used for making optimum decisions in major manufacturing process design, air traffic control, bank teller scheduling, location of fire stations, computer networking, etc. Recently it has been used in the financial risk analysis and is of great help in obtaining optimum decisions of inventory problems.

In this block we have described different methods of generation of random numbers, which are required for generating data from some logical models with random components, for a system under study. The random number generation methods of generating data for discrete variables have been discussed in Unit 13 and the random number generation methods of generating data for continuous variables have been discussed in Unit 14. We also described how to conduct numerical experiments to obtain approximate solutions of some important problems.

Through simulation technique the operation of the model can be studied and from it, properties concerning the behavior of the actual system can be inferred like forest management, epidemics, traffic congestion, etc. In Unit 15, the different steps in setting up simulation are described. The Monte-Carlo simulation technique is also discussed for solving some deterministic and stochastic problem. Many exercises and examples have been given to make you familiar with simulation techniques for solving different types of problems. In Unit 16 we have described various applications of simulations with examples. Different methods of testing the randomness of the generated sequence are also described.

Suggested Readings:

1. Fisher, R. A. and Yates, F. (1963); Statistical Tables (sixth edition), Longman, England.
2. Rand Corporation (1955); A million random digits with 100,000 Normal Deviates, The Free press, Glencoe, III.

Some Further Readings in Simulation:

1. Fishman, G. S. (1995); Monte-Carlo Concepts, Algorithms and Applications, Springer-Verlag, NewYork.
2. Hoover S. V. and Perry, R. F. (1989); Simulation: A Problem Solving Approach, Addison-Wesley, Reading.
3. Law, A. M. and Kelton, W. D. (1991); Simulation Modeling and Analysis (2^{nd} ed.), McGraw Hill, New York.
4. Morgan, B. J. T. (1984); Elements of Simulation, Chapman and Hall, London.
5. Ross, S. M. (2002); Simulation (3rd ed.), Academic Press, London.
6. Rubinstein, R. Y. and Melamed, B. (1997); Modern Simulation and Modeling, Wiley, New York.

Notations and Symbols

IPT	:	Inverse Probability Transformation
PRN	:	Pseudo Random Numbers
U (0,1)	:	Uniform random variable
LCG	:	Linear Congruential Generator
(n) mod m	:	Reminder part of n/m
X	:	Random variable
$F(x) = P(X \leq x)$	:	Cumulative distribution function of X
f(x)	:	Probability density function of X
P(X= x)	:	Probability mass function of X
μ_x	:	Mean of Normal distribution
σ_x	:	Standard deviation of Normal distribution
Z	:	Standard Normal variate
$Be(\alpha, \beta)$	:	Beta distribution
χ^2_m	:	Chi-square variate
$g(x)$	:	Known function of x
$\hat{\theta}$	:	Estimate of θ
θ	:	Deterministic integral function
S_i	:	Service time for i^{th} customer
I_i	:	Inter arrival time between arrival of i^{th} and $(i-1)^{th}$
μ	:	Mean service rate
λ	:	Mean arrival rate
ξ	:	Traffic intensity
W_n	:	System of n^{th} customer
P_{j0}	:	Probability of an observations lying in j^{th} class
H_0	:	Null hypothesis
H_1	:	Alternative hypothesis
n_j	:	Observed numbers of j^{th} category
n_{Pj0}	:	Expected numbers of j^{th} category
D_n	:	Kolmogorov-Smirnov test statistic

UNIT 13 RANDOM NUMBER GENERATION FOR DISCRETE VARIABLES

Structure

13.1 Introduction
Objectives
13.2 Random Numbers and Pseudo Random Numbers (PRN)
13.3 Random Number Generation
Lottery Method
Middle Square Method
Linear Congruential Generator (LCG) Method
Choice of Linear Congruential Generator's
13.4 Inverse Probability Transformation (IPT) Method
IPT Method for Discrete Random Variable
IPT Method for Continuous Random Variable
13.5 Random Number Generation for Discrete Variable
Discrete Uniform Random Variable
Bernoulli Random Variable
Binomial Random Variable
Geometric Random Variable
Negative Binomial Random Variable
Poisson Random Variable
13.6 Summary
13.7 Solutions/Answers

13.1 INTRODUCTION

We shall see in subsequent units that for simulation of any system or process, which contains some random components, we require a method for obtaining random numbers. For example, the queuing and inventory models, which we shall discuss in later units, require inter-arrival times, service times, demand sizes, etc. which are random in nature. For this we may need random variables from Exponential, Gamma distributions, etc. Usually a large number of random numbers from standard uniform distribution U (0, 1) are required which are independently and identically distributed (i.i.d.). Basically, they are required for generation of random variables from non-uniform distribution such as Binomial, Poisson, Normal, etc. They are also required for calculating integral of a function by Monte-Carlo technique. We shall use these random numbers to generate random variables from some popular discrete and continuous distributions, which we shall require for simulation purposes.

The concepts of Random numbers and Pseudo random numbers (PRN) are described in Section 13.2. The different methods of generation of random numbers and pseudo random numbers are explained in Section 13.3. The Inverse probability transformation (IPT) method for generating the discrete and continuous is explored in Section 13.4. In Section 13.5 describes the methods of random number generation for discrete variate from discrete

Uniform, Bernoulli, Binomial, Geometric, Negative Binomial and Poisson distribution.

Objectives

After studying this unit, you will be able to

- define the random numbers and Pseudo random numbers;
- explain Lottery method of generation of random numbers;
- explain Middle Square method of generation of Pseudo random numbers (PRN);
- explain Congruential method of generation of Pseudo random numbers (PRN);
- explain Inverse probability transformation (IPT) method for generating random variables; and
- describe the generation of random numbers from some discrete probability distributions such as Discrete Uniform, Bernoulli, Binomial, Poisson, Geometric, Negative Binoimial, etc.

13.2 RANDOM NUMBERS AND PSEUDO RANDOM NUMBERS (PRN)

The methodology of generating random numbers has a long history. The earliest methods were carried out by hands, throwing dice, dealing out well suffuled cards, etc. Many lotteries are still operated in this way. In the early twentieth century many mechanised devices were built to generate numbers more quickly. Later electronic devices were used to generate random numbers. Rand Corporation (1955) generated a million numbers and are available in table form. An easy method to read from Rand Corporation table or some other statistical tables which contain a large number of random numbers is to select a row or a column of one, two or more figures.

For example, Fisher and Yates (1963) Statistical Tables give 7500 two-figure random numbers arranged in six pages. Suppose one wants to select ten random numbers from 00-99, the simplest way of doing this is to select a row, column or diagonal of two figure numbers randomly and read ten numbers from 00-99 as they appear in a column. If one wants ten uniformly distributed numbers U (0, 1) then one can divide the chosen number by 99 (range of numbers between 00-99). For example, first ten numbers of the first column of the Random Number Table contains two digit numbers as:

03 97 16 12 55 16 84 63 33 57

These numbers are selected randomly (giving equal probabilities), without replacement, from two digit numbers 00-99. If one wants to convert them to uniformly distributed U (0, 1) variables then one has to divide them by 99 which give:

0.030 0.980 0.162 0.121 0.555 0.162 0.848 0.636 0.333 0.575

However, such methods are satisfactory when one requires a very small number. But in most simulations a very large number in thousands and

millions are required. This requires a lot of memory, time and computers are not very efficient. For this purpose we require a procedure which should be fast and does not need much memory. Now modern computer use some inbuilt methods for generating.

Pseudo Random Numbers (PRN) are not random numbers but they behave like random numbers for all practical purposes. That is why they are called Pseudo Random Numbers. One may define PRN as:

A sequence of PRN (U_i) is a deterministic sequence in the interval [0, 1] having the same statistical properties as a sequence of random numbers.

This means that any statistical test applied to a finite part of sequence (U_i), which aims to detect departure from randomness, would not reject the null hypothesis that the sequence consists of random numbers. Hence, we shall generate a deterministic sequence of PRN and then apply some relevant statistical tests to examine the hypothesis that the sequence is random. In case these tests do not reject this hypothesis, we shall take them as random numbers. But this may be noted here that they are not random numbers in the strict sense.

13.3 RANDOM NUMBER GENERATION

In the past, many methods have been used for generation of random numbers but we shall not discuss all of them here and describe only following three methods:

1. Lottery Method
2. Middle Square Method
3. Linear Congruential Method

13.3.1 Lottery Method

This is the simplest method of generating random numbers from U (0,1). In this method, we have ten cards which are made as homogeneous as possible in shape, size, color, etc. and we assign the numbers 0 to 9 on these cards. Then these cards are put in a rotated drum. If we have to draw random numbers of two digits, then we draw a card from the drum and note the number of the drawn card. This card is replaced in the drum and drum is again rotated. Again, we draw a card and note the number of drawn card and this card is replaced in the drum. So we obtain a random number of two digits. Similarly, we can obtain more digital random numbers.

Another way of generating random numbers is to fix up a spinning arrow on a common clock. When the arrow is spin, the number at which it stops would be noted. The arrow is again spin and the number at which it stops would be noted. In this way we find a random number of two digits. Similarly, we can generate another random number of different numbers of digits. Random numbers can also be generated by tossing a coin or dice etc.

Drawback of Lottery Method

The main drawback of this procedure is that if we want to draw large random numbers of digit three or more, then we will have to draw cards or spin the arrow as a large number of times. So this method is much time consuming.

13.3.2 Middle Square Method

Still on a more sophisticated level computers are used for generating the random numbers. With computer it is typically easier to generate the random numbers by an arithmetic process. The method proposed for use on digital computers to generate random number is "Middle Square Method". In this method, if we want to generate n random numbers of r digits then we take a random number of digits r which is generated from any other method, then square this random number. If there are 2r digits in the square then we take the middle r digits as the next random number. If there are less than 2r digits in the square then we put zeroes in front of it to make 2r digits and then take the middle r digits. For example, if we want to generate a four digit integer random number then take a random number (which is generated by any other method) suppose 8937. To obtain the next number in this sequence we square it. We get 79869969 and take the middle four digits 8699. So that the next random number generated is 8699. The next few random numbers in this sequence are 6726, 2390, 7121, …, etc.

If we have to generate a random number of two digits then take a random number (which is generated by any other method) suppose the number is 13, then square of this is 169. The square of this contain (2r - 1) digits, so put a single zero in front of this square to make 2r digits and then take middle two digits as the next random number, so that the next random number generated is 16. The few next random numbers generated in this sequence are 48, 30, 90, …, etc.

Drawbacks of Middle-Square Method

The middle-square method has the following drawbacks:

1. This method tends to degenerate rapidly. A random number may reproduce itself. For example $x_9 = 7600$, $x_{10} = 7600$, $x_{11} = 7600$,…
2. If the number zero is ever generated, all subsequent numbers generated will also have a zero value unless steps are provided to handle this case.
3. A loop many generate i.e. the same sequence of random numbers can repeat. For example $x_{15} = 6100$, $x_{16} = 2100$, $x_{17} = 4100$, $x_{18} = 6100$, $x_{19} = 2100$.
4. This method is slow since many multiplications and divisions are required to access the middle digits in a fixed word binary computer.

13.3.3 Linear Congruential Generator (LCG) Method

A sequence of integers z_1, z_2, ...is generated by a recursive formula

$$z_i = (az_{i-1} + c) \bmod m \qquad \ldots (1)$$

where, m, a and c are positive integers.

Equation (1) can also be written as:

$$z_i = (n) \bmod m$$

by replacing $(az_{i-1} + c)$ by n.

Here **"(n) mod m"** means the remainder part of n/m. Generally "(n) mod m" is always less than m.

For example, if n = 4592 and m = 543

Then "(n) mod m" = "(4592) mod 543" = 248.

It is obvious that all numbers z_i will satisfy;

$$0 \leq z_i \leq m-1$$

Uniformly distributed PRN from U (0, 1) are obtained as u_i's:

$$u_i = z_i/m$$

Whenever, z_i takes the same value as taken earlier in the sequence then the sequence is repeated again. The length of such a cycle is called **period (p)**. It is clear that $p \leq m$. For $p = m$ the LCG is said to be of **full period**.

In actual simulations thousands of random numbers are required and it is desirable to have LCGs of large period, so that the same sequence is not repeated again.

13.3.4 Choice of Linear Congruential Generator's

From the above discussion, it is clear that m should be as large as possible depending on the computer word size. When c = 0 the LCG is called **multiplicative** generator. For a 32-bit computer it has been shown that

$$m = 2^{31}-1, \quad a = 7^5 = 16807, \quad c = 0$$

results in a good multiplicative generator:

$$z_i = (16807\, z_{i-1} + 0) \bmod (2^{31}-1) \qquad \ldots (2)$$

Starting with initial arbitrary seed value z_0, one can generate the required sequence of u_i's.

For example, **s**ome uniform PRN from LCG given in (2) with

$$z_0 = 441, \qquad m = 2^{31}-1$$

$$u_1 = z_1/m = 0.0034, \qquad u_2 = z_2/m = 0.8063$$

E1) Using the following LCG

$x_i = (1573x_{i-1}+19) \bmod (10^3)$

obtain four x_i's with $x_0 = 89$.

E2) Using PRN generated in E1), obtain four U (0, 1) random variables.

13.4 INVERSE PROBABILTY TRANSFORMATION (IPT) METHOD

13.4.1 IPT Method for Discrete Random Variable

Suppose a discrete random variable X takes values at discrete points

$x_1 < x_2 < x_3 < \ldots.$

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i) \qquad \ldots (3)$$

The algorithm for generating a random variable from F(x) is

1 Generate a uniform random variable U (0, 1) using Section 13.3.

2 Determine the smallest integer i for which $u_i \leq F(x_i)$ (using Equation (3)).

3 Take $x = x_i$ which is the desired variable.

Repeat (1)-(3), starting with a new u, each time to obtain more variables.

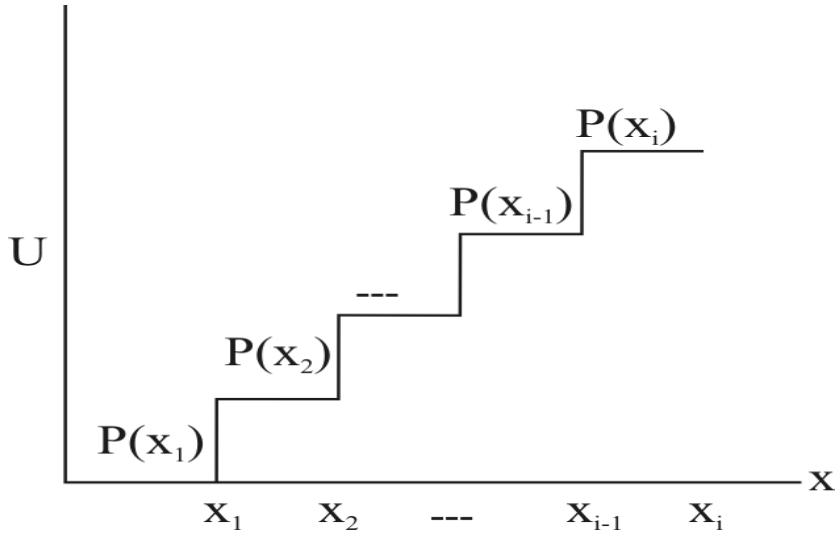


Fig. 1

$$P(X= x_i) = F(x_i) - F(x_{i-1}) = p(x_i) \qquad \text{(See Fig. 1).}$$

Hence, x_i generated in this way has the desired probability $p(x_i)$.

13.4.2 IPT Method for Continuous Random Variable

If F(x) is the cumulative distribution functions of a continuous random variable X,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du$$

where, f(u) is the probability density function (p.d.f.) of the random variable U.

F(x) is strictly increasing function of x. The algorithm is

1 Generate a u~U (0, 1), using Section 13.3.
2 Take F(x) = u and find that x for which F(x) = u. This can be written as $x = F^{-1}(u)$, which is always defined.

X is the desired random variable with distribution F(x).

Proof: We have

$$P(X \leq x) = P (F^{-1}(u) \leq x) = P(u \leq F(x)) = F(x)$$

Which is the desired result (see Fig. 2)

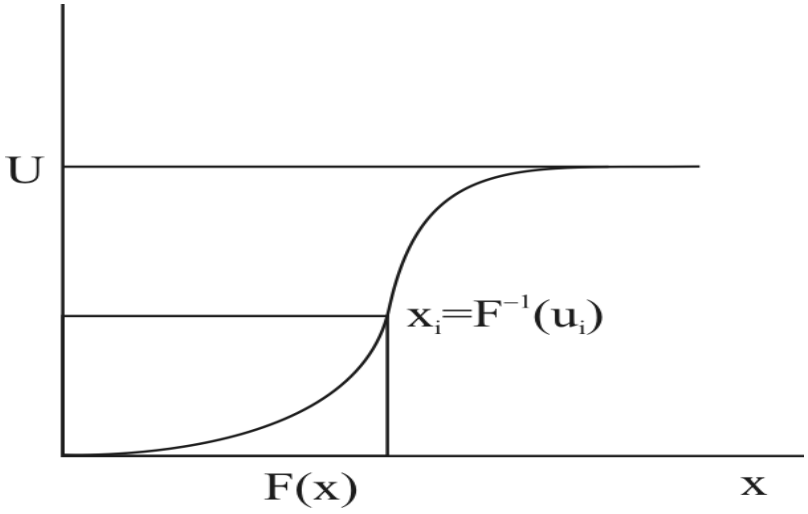


Fig. 2

13.5 RANDOM NUMBER GENERATION FOR DISCRETE VARIABLES

In this section, we shall describe methods for generation of random variables from some important distributions using the PRN described in Section 13.3.

13.5.1 Discrete Uniform Random Variable

Suppose we wish to generate a discrete uniform random variable X whose probability mass function P(x) is given by

$$P(X = i) = 1/N \qquad i = 1, 2, 3, \ldots, N$$

Using Inverse Probabilty Transformation (IPT) method, the algorithm is:

1 Select a uniform random variable u~U (0, 1).

2 Multiply u by N and take x = [Nu] +1

where, [Nu] is the integer part of Nu. X is the desired uniform random variable.

Repeat (1)-(2) to generate more random variables with new u's.

For example, suppose u_1, u_2, u_3 and u_4 are all independent U (0, 1) variables and take values:

$$u_1 = 0.0535, \quad u_2 = 0.5292, \quad u_3 = 0.1189, \quad u_4 = 0.3829$$

Then, we can use them to generate discrete uniform random variables given N = 50. Using the algorithm, we obtain $x_i = [Nu_i] + 1$ and the values are:

$$x_1 = 3, \qquad x_2 = 27, \qquad x_3 = 6, \qquad x_4 = 20$$

13.5.2 Bernoulli Random Variable

Probability mass function of Bernoulli random variable B (1, p) is given by

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

i.e. X takes only two values 1 and 0 with probability p and (1 - p) respectively.

Using IPT method the algorithm is given by

1 Generate a u~U (0, 1).

2 If $u \leq p$ take $x = 1$

$u > p$ take $x = 0$

Repeat (1)-(2) to generate more random variables with new u's.

For example, if we wish to generate three independent Bernoulli random variables for p = 0.3.

And the three independent U (0, 1) are obtained as

$u_1 = 0.928, \quad u_2 = 0.535, \quad u_3 = 0.259$

Using the algorithm we obtain

$x_1 = 0, \quad x_2 = 0, \quad x_3 = 1.$

13.5.3 Binomial Random Variable

Probability mass function of Binomial random variable is given by

$$P(X = x) = {}^{n}C_x \, p^x \, q^{n-x}; \qquad x = 0, 1, 2, \ldots, n.$$

Sum of n independently distributed Bernoulli random variables B (1, p) has a Binomial distribution B (n, p). Therefore, one can generate Binomial random variable by summing n independent Bernoulli random variables, described in Sub-section 13.5.2, algorithm is:

1. Generate n independent B (1, p) described in Sub-section 13.5.2 as $x_1, x_2, \ldots, x_n$.
2. Obtain $x = x_1 + x_2 + \ldots + x_n$

Repeat (1)-(2) with new B (1, p) variables to generate more Binomial random variables.

For example, using four uniform random variables generated in E2), we can generate a binomial random variable B (n, p), with n = 4 and p = 0.3 as follows:

Take $x_i = 1$, if $u_i < 0.3$

$= 0$, otherwise

Therefore, for

$u_1 = 0.016 \quad u_2 = 0.187, \quad u_3 = 0.170, \quad u_4 = 0.429$

we get

$x_1 = 1, \quad x_2 = 1, \quad x_3 = 1, \quad x_4 = 0$

x_i's are independent Bernoulli random variable B (1, 0.3). Sum of n independent Bernoulli random variables has a B (n, 0.3). Hence the value of a Binomial random variable X with n = 4 and p = 0.3 is given by

$$x = x_1 + x_2 + x_3 + x_4 = 3$$

E3) Using five uniform random variables $u_1 = 0.316$, $u_2 = 0.087$, $u_3 = 0.270$, $u_4 = 0.129$, $u_5 = 0.249$, generate a Binomial random variable B (n, p), with n = 5 and p = 0.2.

E4) Using Bernoulli random variables $x_1 = 0$, $x_2 = 1$, $x_3 = 1$, $x_4 = 1$ and $x_5 = 0$ with p = 0.2, generate a Binomial variable with n = 5 and p = 0.2.

13.5.4 Geometric Random Variable

Geometric distribution gives the probability of first success in the n^{th} trial, when the trials are independent and probability of success is p in each trial. The probability is given by

$$P(X = n) = (1-p)^{n-1} p = p\, q^{n-1}, \quad n = 1, 2, 3, \ldots$$

$$P(X \leq t) = p\,(1 + q^1 + q^2 + \ldots + q^{t-1}) = p\,(1 - q^t)/(1 - q)$$

Obtain u~U (0, 1) and using IPT, put $u = P(X \leq t)$. Solving for t gives the following algorithm:

1 Select a u~ U(0,1);

2 t = log (1-u)/log (q), or equivalently t = log(u)/log (q);

where q =1 - p

3 Take x = [t]+1;

where, [t] is the rounded up value of t and log is the simple logarithm at the base e.

Generate more x's by repeating (1)-(3) with new u's.

For example, if four uniform random variables from U (0, 1) are given as:

$u_1 = 0.39$, $u_2 = 0.89$, $u_3 = 0.23$, $u_4 = 0.76$

and if we wish to obtain x_1, x_2, x_3, x_4 which are following geometric distribution with p = 0.3.

So, $q = 1 - 0.3 = 0.7$

Therefore, $t_i = \log(u_i)/\log(q)$ then

$t_1 = 2.64$, $t_2 = 0.33$, $t_3 = 4.12$, $t_4 = 0.76$

$x_1 = 3$, $x_2 = 1$, $x_3 = 5$, $x_4 = 1$

13.5.5 Negative Binomial Random Variable

The distribution of number of failures until k successes is known as Negative Binomial (NB) distribution. It has two parameters p and k, where p is the probability of success in a trial.

$$P(X = x) = {}^{k+x-1}C_{k-1} p^k q^x; \qquad x = 0, 1, 2, \ldots$$

The sum of k independent Geometric variables, each with parameter p, has a NB distribution. Therefore, one can generate Negative Binomial random variable by summing k independent Geometric random variables, described in Sub-section 13.5.4, algorithm is:

1. Generate k independent Geometric random variables with parameter p described in Sub-section 13.5.4 as $x_1, x_2, \ldots, x_k$;
2. Obtain $x = x_1 + x_2 + \ldots + x_n$

Repeat (1)-(2) with new Geometric random variables to generate more Negative Binomial random variables.

For example, if Geometric variables generated in Example given above, for p = 0.3 are

$x_1 = 3, \quad x_2 = 1, \quad x_3 = 5, \quad x_4 = 1$

Therefore, $x = x_1 + x_2 + x_3 + x_4 = 3 + 1 + 5 + 1 = 10$

is the desired value of Negative Binomial variable with k = 4 and p = 0.3

E5) Using Geometric random variables $x_1 = 2$, $x_2 = 3$, $x_3 = 1$, $x_4 = 5$ and $x_5 = 0$ with p = 0.2, generate a Negative Binomial variable with k = 5 and p = 0.2.

13.5.6 Poisson Random Variable

The probability mass function of Poisson random variable with mean λ is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots\ldots$$

IPT method is rather slow for generation of Poisson variables. For this its relationship with Exponential random variable is used. If inter-arrival times are independent and have Exponential distribution with mean 1, then the number of arrivals X in an interval $(0, \lambda)$ time has a Poisson distribution with mean λ.

Suppose $y_1, y_2, y_3, \ldots$ are inter-arrival times for the first, second, third, …….. arrivals and suppose λ lies between n^{th} and $(n+1)^{th}$ arrival, i.e.

$$\sum_{i=1}^{n} y_i < \lambda \leq \sum_{i=1}^{n+1} y_i$$

then, X = n has a Poisson distribution with mean λ. To generate y_i from Exponential distribution with mean 1, one uses

$$y_i = -\log u_i$$

where, $u_i \sim U(0,1)$ and log is simple logarithm at base e. Algorithm is given by:

1. Generate $u_1, u_2, \ldots$ as independent U(0,1)
2. Find cumulative sums of $(n+1)$ $\log u_i$'s and take X = n, when

$$-\sum_{i=1}^{n} \log u_i < \lambda \leq -\sum_{i=1}^{n+1} \log u_i$$

For obtaining more Poisson variables repeat (1)-(2) with new sets of u_i's.

For example, if a sequence of independent u_i's is given by

$u_1 = 0.29, \quad u_2 = 0.89, \quad u_3 = 0.35, \quad u_4 = 0.56, \quad u_5 = 0.69$

So we generate a random variable X which has a Poisson distribution with

$\lambda = 2.5$ as:

$\log u_1 = -1.24$, $\log u_2 = -0.12$, $\log u_3 = -1.05$, $\log u_4 = -0.58$, and

$\log u_5 = -0.37$

Then the cumulative sums of (n+1) $\log u_i$'s

$$-\sum_{i=1}^{3} \log u_i = 2.41 < 2.5 \leq -\sum_{i=1}^{4} \log u_i = 2.99$$

Hence x = 3.

E6) Using uniform random variables generated in E2) generate a Poisson random variable with $\lambda = 4.5$.

E7) Generate a complete cycle for the LCG given below:

$x_i = (5x_{i-1} + 3) \bmod 16, \quad$ with $x_0 = 5$

E8) Generate ten uniform random numbers U(0, 1) from the multiplicative LCG given below:

$x_i = (49x_{i-1}) \bmod 61, \quad$ with $x_0 = 1$

E9) A man tosses an unbiased coin ten times. Using the first ten random numbers generated in E7) obtain a sequence of heads and tails.

E10) Using the first ten random numbers generated in E9) simulate the number of heads obtained in two games of 5 trials each when the probability of obtaining a head is 0.6.

E11) Using uniform random numbers generated in E9) obtain the numbers of trials required for the first success when probability of success p is 0.3.

E12) Using uniform random numbers generated in E9) generate number of trials required for obtaining exactly two successes.

E13) Using the uniform random numbers given below obtain three Poisson random variables when $\lambda = 2$.

0.696, 0.457, 0.493, 0.784, 0.123, 0.478, 0.487, 0.031, 0.681, 0.258

E14) For $\lambda = 2$ cumulative probability for Poisson distribution are given in the following:

X:	0	1	2	3	4

$\sum p(x)$:	0.1353	0.4060	0.6767	0.8570	0.9473
X:	5	6	7	8	9
$\sum p(x)$:	0.9834	0.9955	0.9989	0.9998	1.000

Using uniform random variables given in E13) obtain three Poisson variables with $\lambda=2$, using Inverse Probability Transformation (IPT).

E15) From n = 10 cumulative probability of Binomial distribution with p = 0.25 i.e. B (10, 0.25) is given as:

X	0	1	2	3	4
$\sum p(x)$	0.0563	0.2440	0.5256	0.7759	0.9219
X	5	6	7	≥ 8	
$\sum p(x)$	0.9803	0.9965	0.9996	1	

Using uniform random variables given in E13) obtain three Binomial variables from B (10, 0.25) using IPT.

E16) Suppose a population consists of hundred units numbered as 1, 2, …, 100. One wants to select a sample of five units randomly (with equal probabilities 1/100). Using the uniform random numbers given in E13) select a sample of five units.

13.6 SUMMARY

In this unit, we have discussed:

1. A method for generating Pseudo random numbers (PRN);
2. Methods for generating Uniform random variables U (0,1);
3. Method of Inverse probability transformation(IPT) for generating random variables from a given distribution; and
4. Methods for generating Discrete Uniform, Bernoulli, Binomial, Geometric, Netgative Binoimial and Poisson random variates.

13.7 SOLUTIONS /ANSWERS

E1) We have

$x_i = (1573x_{i-1} + 19) \bmod (10^3)$ with $x_0 = 89$

Therefore,

$x_1 = 140016 \bmod (10^3) = 16$

$x_2 = 25187 \bmod (10^3) = 187$

$x_3 = 294170 \bmod (10^3) = 170$

$x_4 = 267429 \bmod (10^3) = 429$

E2) We have

$m = 10^3$ and $u_i = x_i/m$,

then

$u_1 = 0.016$ $u_2 = 0.187$, $u_3 = 0.170$, $u_4 = 0.429$

E3) Take $x_i = 1$, if $u_i < 0.2$

$= 0$, otherwise

Therefore, for

$u_1 = 0.316$, $u_2 = 0.087$, $u_3 = 0.270$, $u_4 = 0.129$, $u_5 = 0.249$

we get

$x_1 = 0$, $x_2 = 1$, $x_3 = 0$, $x_4 = 1$, $x_5 = 0$

x_i's are independent Bernoulli random variable B (1, 0.2). Sum of n independent Bernoulli random variables has a B (n, 0.2). Hence the value of a Binomial random variable X with n = 5 and p = 0.2 is given by

$x = x_1 + x_2 + x_3 + x_4 + x_5 = 2$

E4) Bernoulli variables for p = 0.2 are given

$x_1 = 0$, $x_2 = 1$, $x_3 = 1$, $x_4 = 1$, $x_5 = 0$

Therefore, $x = x_1 + x_2 + x_3 + x_4 + x_5 = 0 + 1 + 1 + 1 + 0 = 3$

is the desired value of Binomial variable with n = 5 and p = 0.2.

E5) Geometric variables for p = 0.2 are given

$x_1 = 2$, $x_2 = 3$, $x_3 = 1$, $x_4 = 5$, $x_5 = 0$

Therefore, $x = x_1 + x_2 + x_3 + x_4 + x_5 = 2 + 3 + 1 + 5 + 0 = 11$

is the desired value of Negative Binomial variable with k = 5 and p = 0.2.

E6) We have

$u_1 = 0.016$, $u_2 = 0.187$, $u_3 = 0.170$, $u_4 = 0.429$

Therefore,

$\log u_1 = -4.13$, $\log u_2 = -1.68$, $\log u_3 = -1.77$, $\log u_4 = -0.85$

$-\log u_1 = 4.13 \leq 4.5 \leq -\log u_1 - \log u_2 = 5.81$

Hence, Poisson random variable X is given by x =1

E7) A full cycle of random numbers generated from LCG

$$x_i = (5x_{i-1} + 3) \bmod 16, \quad \text{with } x_0 = 5$$

is given as follows:

x: 12, 15, 14, 9, 0, 3, 2, 13, 4, 7, 6, 1, 8, 11, 10, 5

E8) The given LCG is

$$x_i = (49\ x_{i-1}) \bmod 61, \quad \text{with } x_0 = 1$$

and we also have $u_i = x_i / 61$

x :49	22	41	57	48	34	19	16	52	47
u :0.80	0.36	0.67	0.93	0.79	0.56	0.31	0.26	0.85	0.77

E9) We have given LCG

$$x_i = (5x_{i-1} + 3) \bmod 16, \quad \text{with } x_0 = 5$$

which gives 10 random numbers as

12, 15, 14, 09, 00, 03, 02, 13, 04, 07

We also have $u_i = x_i /16$

Therefore,

u:	0.75	0.94	0.87	0.56	0.00	0.19	0.12	0.81	0.25	0.44
x :	H	H	H	H	T	T	T	H	T	T

(Taking $u \geq 0.5$ as Head (H) and $u < 0.5$ as Tail (T))

E10) This is the case of Binomial, B (5, 0.6) using the uniform random numbers generated in E9) take H when $U \leq 0.6$ and T when $U > 0.6$, and sum the number of the heads in first five trials and the next five trials. The numbers are:

x: 2, 4

E11) This is the case of Geometric distribution. We have

$$t = \log (u)/\log (q)$$

and $x = [t] + 1$

$$t = \log(u_1)/\log (q)$$

$$= \log (0.75)/\log (0.7)$$

$$= -0.2877/-0.3567$$

$$= 0.80$$

Therefore, $x_1 = [0.80] + 1 = 1$

Hence $x_1 = 1$

E12) The random variable required is Negative Binomial with p = 0.3 and k = 2. As in E1), we obtain the next Geometric variable by taking:

$$t = \log(u_2)/\log(q)$$

$$= \log(0.94)/\log(0.7)$$

$$= -0.0619/-0.3567$$

$$= 0.17$$

Therefore $x_2 = [0.17] + 1 = 1$

Hence $x_2 = 1$.

The Negative Binomial variable is the sum of k independent Geometric random variables, and thus

$$x = x_1 + x_2 = 1 + 1 = 2$$

E13) We have

S.No.	1	2	3	4	5	6	7	8	9	10
u_i	0.696	0.457	0.493	0.784	0.123	0.478	0.487	0.031	0.681	0.258
$-\log u_i$	0.36	0.78	0.70	0.24	2.09	0.73	0.71	3.47	0.38	1.35

Therefore,

$$-\sum_{i=1}^{3}\log u_i = 1.85 < 2 \leq -\sum_{i=1}^{4}\log u_i = 2.09$$

Hence, $x_1 = 3$.

Starting from u_5

$$-\log u_5 > 2$$

Hence, $x_2 = 0$.

Starting from u_6

$$-\sum_{i=6}^{7}\log u_i = 1.457 < 2 \leq -\sum_{i=6}^{8}\log u_i = 4.931$$

Hence, $x_3 = 2$

Thus, three Poisson random variables generated for $\lambda = 2$ are:

3, 0, 2

E14) We have $u_1 = 0.696$, $u_2 = 0.457$, $u_3 = 0.493$

Using IPT, we see that

$P(X \leq 2) < u_1 = 0.696 \leq P(X \leq 3)$, $x_1 = 3$

$P(X \leq 1) < u_2 = 0.457 \leq P(X \leq 2)$, $x_2 = 2$

$P(X \leq 1) < u_3 = 0.493 \leq P(X \leq 2)$, $x_3 = 2$

Hence three random variables from P (2) are:

3, 2, 2.

E15) We have $u_1 = 0.696$, $u_2 = 0.457$, $u_3 = 0.493$.

$P(X \leq 2) < u_1 = 0.696 \leq P(X \leq 3)$, $x_1 = 3$

$P(X \leq 1) < u_2 = 0.457 < P(X \leq 2)$, $x_2 = 2$

$P(X \leq 1) < u_3 = 0.493 \leq P(X \leq 3)$, $x_3 = 2$

Hence three random variables from B (10, 0.25) are:

$x_1 = 3$, $x_2 = 2$, $x_3 = 2$.

E16) This is the case of discrete uniform random variable with N =100. Using the random numbers in E13) we obtain the following numbers:

$x_i = [Nu_i] + 1$

x : 70, 46, 50, 79, 13.

UNIT 14 RANDOM NUMBER GENERATION FOR CONTINUOUS VARIABLES

Structure

14.1 Introduction
Objectives
14.2 Uniform Distribution
14.3 Exponential Distribution
14.4 Normal Distribution
Inverse Probability Transformation Method
Central Limit Theorem Approximation
Box-Muller Method
Marsaglia's Polar Method
14.5 Gamma, Chi-square and Beta Distributions
14.6 Poisson Process
14.7 Summary
14.8 Solutions/Answers

14.1 INTRODUCTION

In Unit 13, we have given some methods of generating variates from some important discrete distributions. In this unit we shall describe generation of variates from some important continuous distributions. Sometimes we shall use the IPT method for generation of variates from a particular distribution but in some cases we shall use some other methods which are more convenient and efficient.

Method of generation of random variates from continuous Uniform and Exponential probability distribution is explained in Sections 14.2 and 14.3. In Section 14.4, the some useful methods of generation of random variates from Normal distribution i.e. Inverse Probability Transformation method, Central Limit Theorem approximation method, Box-Muller method and Marsaglia's Polar method are explored with examples. In Section 14.5, the methods of generating the variates from the Gamma, Chi-square and Beta distributions are explained whereas a brief discussion has been done about the Poisson process and Waiting Time distributions in Section 14.6.

Objectives

After studying this unit, you would be able to

- describe the generation of variables from continuous distributions;
- explain the method of generation of variables from Uniform distribution;
- explain the method of generation of variables from Exponential distribution;
- explain the method of generation of variables from Normal distribution;

- describe the method of generation of variables from Gamma distributions; and
- describe the method of generation of random variates from Poisson Process.

14.2 UNIFORM DISTRIBUTION

We have already described generation of Uniform random variable U(0,1) in Sub-section 13.3.5. Now if one wishes to generate variates from U(a,b) whose probability density function (p.d.f.), f(x), is given by

$$f(x) = \begin{cases} \dfrac{1}{(b-a)}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$$

A simple transformation of u~ U (0, 1) will convert it to x ~ U (a, b)

$$u = (x - a)/ (b - a)$$

Hence $x = a + (b - a)\, u$

Following is the algorithm:

1. Generate a uniform variate u ~ U(0,1)
2. Take $x = a + (b - a)\, u$

Repeat the steps (1) - (2) with a new u to generate more variates.

14.3 EXPONENTIAL DISTRIBUTION

The probability density function of Exponential distribution with parameter α is given by

$$f(x) = \alpha e^{-\alpha x}, \qquad \alpha > 0,\ x \geq 0$$

where, 1/ α is the mean of the distribution.

Distribution function F(x) of Exponential distribution is given by

$$F(x) = 1 - e^{-\alpha x}$$

Using IPT described in Sub-section 13.4.2 of Unit 13 and equating u~ U (0, 1) to F(x)

$$u = F(x) = 1 - e^{-\alpha x}$$

giving

$$x = -[\log (1 - u)]/ \alpha$$

or equivalently

$$x = -[\log (u)]/ \alpha$$

Algorithm is

1. Generate a uniform variate u ~ U (0, 1)
2. Take x = − [log (u)]/α, which has Exponential distribution with parameter α.

For example, if u = 0.5921. To generate an exponential variate with α = 2, take

$$x = -\log(u)/\alpha = 0.262$$

x is the desired exponential variate with α = 2.

14.4 NORMAL DISTRIBUTION

The probability density function of Normal distribution N (μ, σ^2) is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\ \sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad , -\infty < x < \infty$$

Transformed variable

$$z = \frac{(x-\mu)}{\sigma} \text{ is following N (0, 1)}$$

The cumulative distribution function (c.d.f.), of z denoted by Φ (z), is given by

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

14.4.1 Inverse Probability Transformation Method

We generate a u ~ U (0,1) and using inverse probability transformation, we equate

$$u = \Phi(z)$$

$$z = \Phi^{-1}(u)$$

Then we obtain z from standard normal cumulative probability table. Then X~ N (μ, σ^2) is given by equating

$$z = \frac{(x-\mu)}{\sigma}$$

giving, $x = \mu + \sigma z$

For example, if we wish to generate a X~ N (μ, σ^2), with μ = 5.0 and

σ^2 = 2.0.

Suppose, the generated u~U (0, 1) is 0.69.

Equating u to Φ (z), we obtain

$$\Phi(z) = 0.69$$

Using the normal c.d.f. table, we have

$$z = \Phi^{-1}(0.69) = 0.50$$

$$z = \frac{(x - 5.0)}{\sqrt{2.0}}$$

giving $x = 5.0 + \sqrt{2.0}\ (0.50)$

$$= 5.707$$

Hence, x is the desired normal variate.

As the integral Φ (z) is not in a closed form, one has to use normal probability table which is not very convenient when one has to generate a very large number of variates. We shall use an alternative method, which is more popular and efficient for generation of normal variates.

14.4.2 Central Limit Theorem Approximation

The slow but simple approach makes use of the Central limit theorem (CLT) from mathematical statistics. An often overlooked consequence of the CLT is its assertion that the sum of n independent samples from identical distributions, such as $x_1 + x_2 + ...+ x_n$, is normally distributed with the mean equal to $n\mu_x$ and a variance of $n\sigma_x^2$ where μ_x and σ_x^2 are mean and variance of the variable X respectively. If the distribution of X is uniform on the (0, 1) interval, then a sum $y = x_1 + x_2 + ... + x_n$ of n such numbers is normally distributed with

$$\mu_y = n/2 \text{ and } \sigma_y^2 = n/12$$

so that

$$z = \frac{y - n/2}{\sqrt{n/12}}$$

is standard normal with $\mu_z = 0$ and $\sigma_z^2 = 1$.

The CLT informs us that the larger the value of n, the better the approximation to the Normal distribution. For most purposes, a value of 12 is convenient and leads to a reasonable approximation. In this case, z equation will be

$$z = \frac{y - 12/2}{\sqrt{12/12}} = y - 6$$

The CLT generator is not the best one if more than a few numbers are to be generated. It is very slow and it does not adequately sample the extreme tails of the Normal distribution.

14.4.3 Box-Muller Method

This method is very popular and used often to generate normal random variable. We shall give only algorithm here but for details see Ross (2002).

Algorithm

1. Generate two independent uniform U(0,1) variates u_1 and u_2.
2. $x = \sqrt{(-2\log u_1).\cos(2\pi u_2)}$

   $y = \sqrt{(-2\log u_1).\sin(2\pi u_2)}$

   where, $2\pi = 360^o$.

This method gives two normal independent N (0,1) variables x and y corresponding to two independent $u_1$and u_2. x and y thus generated are desired N (0, 1) variables. They have to be transformed to obtain N (μ, σ^2). More variables can be generated by repeating the steps (1)-(2) with new independent u_i's.

14.4.4 Marsaglia's Polar Method

Complexity arises in generating Sines and Cosines of random angles in using above method. Marsaglia (1962), gave the following modification in the Box-Muller transformation that avoids evaluation of the trigonometric function:

1. Generate u_1 and u_2 from U (0, 1) and set $v_1 = 2\,u_1 - 1$ and $v_2 = 2\,Uu_2 - 1$
2. If $v_1^2 + v_2^2 > 1$ go to step 1
3. Deliver

   $$C = \left[\frac{-2\,\mathrm{Ln}\left(v_1^2 + v_2^2\right)}{\left(v_1^2 + v_2^2\right)}\right]^{\frac{1}{2}}$$

   where, Ln is the natural logarithm at the base e.
4. $x = Cv_1$, $y = Cv_2$

E1) Six independent $u_i \sim U(0,1)$ are given below:

$u_1 = 0.59$, $u_2 = 0.32$, $u_3 = 0.89$, $u_4 = 0.17$, $u_5 = 0.92$, $u_6 = 0.66$

Using Box-Muller method, generate 6 independent N (3, 1.5) variates.

14.5 GAMMA, CHI-SQUARE AND BETA DISTRIBUTIONS

14.5.1 Gamma Distribution

The probability distribution function of Gamma distribution G(n, α) is given by

$$f(x) = \frac{\alpha^n x^{n-1} e^{-\alpha x}}{\Gamma(n)}, \qquad 0 \le x < \infty,\ \alpha > 0,\ n > 0$$

The cumulative distribution function is not in closed form and thus IPT method cannot be applied. Therefore, some alternative methods have been used.

In case where n is an integer then G (n, α) can be obtained by summing n exponential variables with parameter α.

Thus, the algorithm is:

1. Generate n independent uniform U (0,1) variables $u_1, u_2, \ldots, u_n$.
2. Obtain n independent exponential variates y_i's with parameter α.

$$y_i = -\log u_i/\alpha, \qquad i = 1, 2, \ldots, n$$

3. $x = \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{-\log u_i}{\alpha}$

This is also known as **Erlang** distribution.

When n is not an integer then the generation of Gamma variable is rather complicated. For generation of variables for non-integer values of n see Law and Kelton (1982).

For example, if we wish to generate two random variates from G(2, 0.5) and suppose we have generated two independent u_i's as

$$u_1 = 0.49, \qquad u_2 = 0.83,$$

Therefore, compute

$$x = \sum_{i=1}^{2} \frac{-\log u_i}{0.5} = 1.799$$

14.5.2 Chi-square Distribution

Chi-square distribution, denoted by χ^2_m, is a particular case of Gamma distribution with n = m/2, where m is even and α = 1/2. The probability distribution function is

$$f(x) = \frac{x^{(m/2)-1} e^{-x/2}}{2^{m/2}\Gamma(m/2)}, \qquad x > 0$$

Chi-square variate χ^2_m may also be defined as sum of square of m independent standard normal variates, so it can be generated by the sum of squares of m independent N (0, 1) variables. Algorithm is:

1. Generate m independent N (0,1) variates z_i's, (i = 1, 2,…, m)
2. Take $x = \sum_{i=1}^{m} z_i^2$
3. x is the desired χ^2_m variate on m degrees of freedom.

14.5.3 Beta Distribution

A random variable X has a Beta distribution if the probability distribution function is given by:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \; ; \; \alpha > 0, \beta > 0, 0 \le x \le 1$$

and is denoted by Be (α, β).

An easy method of generation of Beta random variate is the following:

If two variates y_1 and y_2 are independently distributed as G(α,1) , G(β,1), respectively, then

$$x = y_1/ (y_1+y_2)$$

has a Be(α, β) distribution.

E2) Using first four N (0, 1) variates generated in E1) obtain a χ^2 variate with four degrees of freedom.

14.6 POISSON PROCESS

A random variable which is often used for simulation in queuing theory is known to be distributed as Poisson Process.

If the inter-arrival times $x_1, x_2, \ldots$ are independent and identically distributed as exponential random variables with parameter α, then the times at which events occur is known to have a poisson process. In this case, number of events occurring in an interval (0, t) has a Poisson distribution with mean αt. Seeing the above relationship between Exponential distribution and poisson process one can easily generate poisson process variate.

Algorithm is given as follows:

1. Generate $u_1, u_2, \ldots$ from U (0, 1).
2. Find exponential variates y's by taking

   $y_i = -(\log u_i)/ \alpha, i = 1, 2, \ldots$
3. Take $t_1 = -(\log u_1) /\alpha = y_1$

   $t_2 = t_1 - (\log u_2) / \alpha = y_1+y_2$

   $t_3 = t_2 - (\log u_3) / \alpha = y_1+y_2+y_3$

   .

   .

   .

   $t_i = t_{i-1} - (\log u_i) / \alpha = y_1+y_2+\ldots+y_i$

$t_1, t_2, t_3, \ldots$ are the times at which event occur and follow poisson process with rate α.

For example, if six U (0, 1) variables u_i's are given as:

$u_1 = 0.78, u_2 = 0.31, u_3 = 0.76, u_4 = 0.23, u_5 = 0.79, u_6 = 0.96$

so we generate timing (t_i) of event occurring from a poisson process when rate of events occurring per unit time is 0.3 i.e. $\alpha = 0.3$.

Then, the calculated $y_i = -(\log u_i)/ \alpha$ are:

$y_1 = 0.828, y_2 = 3.904, y_3 = 0.915, y_4 = 4.899, y_5 = 0.786, y_6 = 0.136$

Therefore, $t_i = \sum_{j=1}^{i} y_j$

$t_1 = 0.828$, $t_2 = 4.732$, $t_3 = 5.647$, $t_4 = 10.565$, $t_5 = 11.332$, $t_6 = 11.468$

t_i's are the required timings at which the events occur.

E3) The inter-arrival times of patients arriving in a clinic has an Exponential distribution with rate $\alpha = 0.2$ per minute. Simulate the times of six patients arriving in the clinic. Also give the number of patients arriving in first 20 minutes.

(Hint: Use LCG $x_i = (1573\ x_{i-1} + 19) \bmod 10^3$, starting with $x_0 = 159$.)

E4) A train is expected to arrive in a station at 8:00 AM. However, it has been observed that it reaches station between 7:55 A.M. to 8:05 AM and the times are uniformly distributed between the above intervals. Using the following U (0, 1) random numbers simulate time for arrival on ten days:

U (0, 1): 0.579 0.052 0.312 0.307 0.645 0.945 0.645 0.956
0.394 0.110 0.854 0.476

E5) In a warehouse three trucks arrive per hour to be unloaded. Generate inter-arrival times of ten trucks using random numbers of E4), assuming that distribution of inter-arrival times is exponential. In how many cases it exceeds half an hour?

E6) The mileage (in thousand of miles) which car owners get with a certain kind of radial tyre is a random variable having an Exponential distribution with $\alpha = 0.025$. Generate the mileage of five such tyres (use random numbers of E4). Obtain the average mileage of these five tyres.

E7) The probability density of X is given by

$$f(x) = \begin{cases} \frac{x}{2} & ; \quad \text{for } 0 < x < 2 \\ 0 & , \quad \text{elsewhere} \end{cases}$$

Generate five random variables from f(x) using U (0, 1) from E4).

E8) If u_i's are independent U (0, 1) random variables then define y as sum of n "u_i's" as

$$y = \sum_{i=1}^{n} u_i$$

$E(y) = n/2$, $V(y) = n/12$

If n is large then $z = [y - E(y)]/\sqrt{V(y)}$

is approximately distributed as N(0, 1). Using ten u_i's given in E4) generate a normal random variate N (0, 1) in this way.

E9) Give an algorithm using IPT method to generate variate from the following Beta probability density function:

$$f(x) = 6x(1 - x), \qquad 0 \le x \le 1$$

Find x when u = 0.5.

E10) Distribution function of Pareto random variable is given by

$$F(x) = 1 - \left(\frac{k}{x}\right)^a, \qquad a > 0,\ 0 < k \leq x$$

Given a u~U (0, 1) generate x, when a = 2 and k = 1. Suppose u = 0.5 then find x.

E11) If x and y are distributed normally and independently as N (2, 2) and N (3, 5), respectively. Define a new variate z = x + y. Generate a new variable z using uniform variates given in E4).

E12) Describe an algorithm using IPT for generating a variate from Logistic distribution whose p.d.f. is given by:

$$f(x) = \frac{e^{-x}}{\left(1+e^{-x}\right)^2}, \qquad -\infty < x < \infty$$

Take u = 0.3 to generate a variate x from the above distribution.

E13) In a certain city, the daily consumption of water (in millions of litres), follows approximately a Gamma distribution, G (n, α) with n = 3 and α = 1/3. Using random numbers given in E4) generate daily consumption of water for three days from this distribution.

E14) If the annual proportion of new restaurants that fail in a given city may be looked upon as a random variable having a Beta distribution with α = 1 and β = 4, then generate proportion of restaurants that are expected to fail in one year, using random numbers given in E4).

E15) The life in years of a certain type of electrical switch has an Exponential distribution with α = 0.5. Generate life time (in years) for ten switches using uniform variates in E4). What is the proportion of bulbs with life more than one year in the sample obtained?

E16) Buses arrive at a sporting event according to Poisson process with rate of five per hour. Write an algorithm to simulate the arrival of buses by time t =1, using uniform variable in E4)

14.7 SUMMARY

In this unit, we have discussed:

1 Method for generation of random variables from Uniform distribution;
2 Method for generation of random variables from Exponential distribution;
3 Various methods for generation of random variables from Normal distribution;
4 Methods for generation of random variables from Gamma distribution;
5 Methods for generation of random variables from Chi-square distribution;
6 Methods for generation of random variables from Beta distribution; and
7 Poisson process which is useful in simulation of queuing problems.

14.8 SOLUTIONS / ANSWERS

E1) Taking pairs (u_1, u_2), (u_3, u_4), (u_5, u_6) and using Box-Muller transformation we obtain the following N (0, 1) variates:

$y_1 = -0.249$, $y_2 = 0.999$, $y_3 = 0.276$, $y_4 = 0.396$, $y_5 = -0.339$,

$y_6 = -0.227$

Corresponding N (3,1.5) variates are obtained by

$x_i = (\sqrt{1.5})\, y_i + 3.0$ giving

$x_1 = 2.695$, $x_2 = 4.223$, $x_3 = 3.338$, $x_4 = 3.485$, $x_5 = 2.585$, $x_6 = 2.722$

E2) N (0,1) variables y_i's generated in E1) are given as

$y_1 = -0.249$, $y_2 = 0.999$, $y_3 = 0.276$, $y_4 = 0.396$

χ^2_4 variable $x = \sum_{i=1}^{4} y_i^2 = 1.293$

E3) We have $u_1 = 0.126$, $u_2 = 0.217$, $u_3 = 0.577$, $u_4 = 0.640$, $u_5 = 0.739$,

$u_6 = 0.466$ then

$y_i = -(\log u_i)/\alpha, \qquad i = 1, 2, \ldots, 6$

$y_1 = 10.357$, $y_2 = 7.639$, $y_3 = 2.749$, $y_4 = 2.231$, $y_5 = 1.512$,

$y_6 = 3.818$

Therefore, $t_i = \sum_{j=1}^{i} y_j$

$t_1 = 10.357$, $t_2 = 17.996$, $t_3 = 20.745$, $t_4 = 22.976$, $t_5 = 24.488$,

$t_6 = 28.306$.

Number of patients in first twenty minutes = 2

E4) This is the case of uniform distribution U (a, b) with a = 7:55,

b = 8:05, therefore

$x = 7{:}55 + 10\, u$

Using u's, we get x's as:

8:01, 7:55, 7:58, 7:58, 8:01, 8:04, 7:59, 7:56, 8:04, 8:00.

E5) Exponential variate is given by

$x_i = -(\log u_i)/\alpha = -(\log u_i)/3$

This gives x_i as

x: 0.182, 0.985, 0.388, 0.394, 0.146, 0.015, 0.310, 0.736, 0.052, 0.247

(where x is in hours)

In two cases it exceeds half an hour.

E6) We have $x_i = -(\log u_i)/\alpha = -(\log u_i)/0.025$

x: 21.86, 118.26, 46.59, 47.24, 17.54

Average = 50.30 (in thousand miles).

E7) Cumulative distribution function F(x) is given by

$$F(x) = \int_0^x \frac{u}{2}\, du$$

$$F(x) = \frac{x^2}{4}, \qquad 0 < x < 2$$

Equating uniform U (0, 1) random variable u, we have

$$u = \frac{x^2}{4} \Rightarrow x = 2\sqrt{u}$$

This gives 1.522, 0.456, 1.117, 1.108, 1.606.

E8) We have $y = \sum_{i=1}^{10} u_i = 4.685$

Then E (y) = n/2 = 10 / 2 = 5.0;

V(y) = n / 12 = 10/12 = 0.833

Therefore,

$z = (4.685 - 5.0) / \sqrt{0.833}$

$= -0.315 / 0.913$

$= -0.345$

z is the required N (0,1) variate.

E9) We have $f(x) = 6x(1-x) \qquad 0 \le x \le 1$

$$F(x) = \int_0^x 6u(1-u)\, du$$

$= 3x^2 - 2x^3 \qquad 0 \le x \le 1$

Generate a u ~ U (0, 1), Put

$$u = 3x^2 - 2x^3$$

and obtain $0 \leq x \leq 1$ which satisfies the above equation. This gives a random variate from f(x).

For u = 0.5 the value of x = 0.50.

E10) Equate

$$u = 1 - \left(\frac{k}{x}\right)^a$$

$$x = \frac{k}{(1-u)^{1/a}}$$

when u = 0.5 then

$$x = \frac{1}{(1-0.5)^{1/2}} = \frac{1}{\sqrt{0.5}}$$

$$= 1/0.707 = 1.414$$

Hence, x = 1.414 is the generated variate.

E11) Form E4), we have, $u_1 = 0.579$, $u_2 = 0.052$

Using Box-Muller transformation

$x_1 = \sqrt{(-2 \log u_1).\cos(2\pi u_2)} = 1.045 \times 0.957 = 1.000$

$x_2 = \sqrt{(-2 \log u_1).\sin(2\pi u_2)} = 1.045 \times 0.290 = 0.303$

If x, y are two independent standard normal variates then we know that

$$z = x + y$$

is also normally distributed with mean as

$$E(z) = E(x) + E(y)$$

$$= 2 + 3 = 5,$$

and variance V(z) as

$$V(z) = V(x) + V(y)$$

$$= 2 + 5 = 7$$

Now, z can be generated from x_1 by taking

$$x_1 = \frac{z-5}{\sqrt{7}}$$

giving $z = 5 + \sqrt{7}\, x_1 = 7.645$, which is the desired variate.

E12) Cumulative density function for a Logistic distribution is given by

$$F(x) = \frac{1}{\left(1 + e^{-x}\right)}$$

Suppose u is U (0, 1) random variate then taking

$$u = \frac{1}{\left(1 + e^{-x}\right)}$$

giving

$$x = -\log [(1-u) /u]$$

Put u = 0.3

$$x = -\log [0.7/0.3]$$

$$= -\log (2.333) = -0.847$$

Hence, the logistic variate generated is -0.847.

E13) If n is an integer then we can generate a Gamma (n, α) variate by summing n exponential variate with parameter α. Hence the algorithm is

$$x = -\left(\sum_{i=1}^{n} \log u_i\right) / \alpha$$

Using u_i's in E4) we obtain Gamma variate G (3, 1/3)

$$n = 3, \alpha = 1/3$$

$$x_1 = 3\sum_{i=1}^{3} \log u_i \ , \ x_2 = 3\sum_{i=4}^{6} \log u_i \ , \ x_3 = 3\sum_{i=7}^{9} \log u_i$$

$x_1 = 13.971$, $x_2 = 8.445$ and $x_3 = 4.242$

E14) Using uniform random numbers given in E4) we generate

$y_1 \sim G(1, 1)$ and $y_2 \sim G(4, 1)$ and then take $x = y_1 / (y_1 + y_2)$.

Take

$$y_1 = -\log u_1/1, \quad \text{and} \quad y_2 = -\sum_{i=2}^{5} \log u_i /1$$

$y_1 = 0.546$, and $y_2 = 5.302$

$x = 0.546/5.848 = 0.093$

Hence one Beta (1, 4) variate generated is 0.093.

E15) Exponential variate with parameters α can be generated from uniform random variables U as

$$x = -[\log u] / \alpha$$

using u's given in E4) we have the following:

1.09, 5.91, 2.33, 2.36, 0.88, 0.09, 1.86, 4.41, 0.32, 1.48

Proportion of bulbs with life more than one year is $\frac{7}{10} = 0.7$.

E16) Generate times of arrival of buses by Poisson process with $\alpha = 5$. Using u_i's from E4) we have times of arrival of buses t_i as

$$t_1 = -[\log u_1] / \alpha = 0.109$$

$$t_2 = t_1 - [\log u_2] / \alpha = 0.700$$

$$t_3 = t_2 - [\log u_3] / \alpha = 0.933 \quad .$$

$$t_4 = t_3 - [\log u_4] / \alpha = 1.169$$

Hence from this simulation three buses arrive in the first one hour.

UNIT 16 APPLICATIONS OF SIMULATION

Structure

16.1 Introduction
Objectives
16.2 Some Applications of Simulation
A Waiting Time Model (Queuing Model)
Simulation of a Very Simple Inventory Problem: An Example
16.3 Test for Randomness
Chi-Square Goodness of Fit Test
Kolmogrov-Smirnov Goodness of Fit Test
Runs Up and Down and Test (Test of Independence)
16.4 Advantages and Disadvantages of Simulation
16.5 Summary
16.6 Solutions / Answers

16.1 INTRODUCTION

In Unit 15, we have discussed the preliminaries of starting simulation and general problems that can be solved by simulation. In this unit, we shall consider some important applications of simulations. In this unit, we shall be discussing the applications of simulation in the area of Queuing theory and Inventory problems. Some other problems are also tackled with simulation i.e Air traffic problem control, Bank teller scheduling, Fire station location and Computer network problems. We shall discuss advantages and disadvantages of solving problem by simulation.

Some of the important applications i.e A waiting time model (Queuing Model) and simulation of a very simple inventory problem are explored in Section 16.2. Tests of randomness i.e. Chi-square goodness of fit test, Kolmogrov-Smirnov goodness of fit test and Runs up and down test (Test of independence) are discussed in Section 16.3. The advantages and disadvantages of simulation technique are described in Section 16.2.

Objectives

After studying this unit, you would be able to:

- describe certain applications of simulation with examples i.e. in inventory prblems, in queening theory, etc;
- describe advantages and disadvantages of simulation;
- tests the randomness of the generated sequence i.e. Chi-square goodness of fit test, Kolomogrov-Smirnov goodness of fit test and Runs up and down test of independence; and
- explain the methods of testing the randomness of the generated sequence.

16.2 SOME APPLICATIONS OF SIMULATION

Simulation is becoming a very popular tool for the analysis of a wide variety of the problems. In the last few years it has been used in all major manufacturing process design. Some problems that can be tackled by simulation are:

1. Capacity and feasibility
2. Comparing alternatives
3. Troubleshooting fine tuning

Some problems that can be tackled in other areas are:

- Air traffic control (delays in landing);
- Bank teller scheduling (customer waiting times);
- Location of fire stations (response times);
- Computer network (delays); etc.

Variables given within brackets are performance measures.

Now-a-days a lot of statistical inference methods are also based on Monte-Carlo simulation techniques such as Jack-Knife, Bootstrap, Markov Chain Monte-Carlo (MCMC) methods.

In the following sections, we shall give examples of some important applications.

16.2.1 A Waiting Time Model (Queuing Model)

These models can be used in a wide variety of situations in the real world, where entities (or customers) enter a system for some service. The time between two customers arrival is taken as random variable. The time of serving of each customer is also taken as random variable. Examples of queuing theory are many as follows:

1. Patients arriving at an emergency room and waiting for doctor;
2. Job on a computer centre waiting for CPU time;
3. Airplanes entering an airport's airspace waiting for an available runway; etc.

One may like to use these models to study

(i) What is the average amount of time a customers or job spends for waiting for service?

(ii) What is the probability that the number of customers or job in the system will exceed some fixed level?

(iii) Total number of customers that have been served in an interval [0, t].

If the arriving customer finds the server free, his service starts immediately, and he departs from the system after completion of his service. If the arriving customer finds the server busy, he waits till his turn to be served. Customers are served on first-come-first-out basis (FCFO).

Let S_i denote the service time of the i^{th} customer who arrives at time t_i and let

$$I_i = t_i - t_{i-1}, \quad i \geq 1$$

denote the inter-arrival time between arrival of the i^{th} and $(i-1)^{th}$ customers. Assume that S_i $(i \geq 1)$ and I_i $(i \geq 1)$, each are independent random variables and follow exponential distribution.

Let μ and λ be the mean service rate (customers/unit of time) and mean arrival rates (customers/unit of time), respectively.

Then by inverse probability transformation (IPT) method

$$F(X) = P[X \leq x] = u$$

$$\Rightarrow 1 - e^{-I\lambda} = u$$

$$\Rightarrow e^{-I\lambda} = 1 - u \approx u$$

$$\Rightarrow -I\lambda = \log u$$

$$\Rightarrow I = (-\log u) / \lambda$$

Similarly,

$$S = -(\log u) / \mu$$

The parameter

$$\rho = \lambda / \mu$$

is known as **traffic intensity** and measures the congestion of the system.

If $\rho < 1$, then the system is to be stable.

Note that the total time spent in the system of the n^{th} customer, W_n, from arrival till departure, may be simply expressed as

$$W_n = W_{n-1} - I_n + S_n \qquad \text{if } W_{n-1} > I_n,\ n = 1, 2, \ldots$$

$$= S_n \qquad \text{if } W_{n-1} < I_n$$

starting with $W_1 = S_1$

For general random variables S_i and I_i and single server (GI/G/I) queuing model, it is very difficult to find mean value of W_i, i.e. $E(W_i)$, analytically and simulation may be used. In order to estimate E (W) we run the queuing system N times, each time starting from $W_1 = S_1$.

Then we obtain a sequence of service times $\{S_{ik}, i \geq 1, k = 1, 2, \ldots, N\}$ and sequence of inter-arrival times $[I_{ik}, i \geq 1, k = 1, 2, \ldots, N]$ and calculate W_{ik} and estimate $E(W_i)$ by the sample mean

$$\overline{W} = \sum_{k=1}^{N} \frac{W_{ik}}{N}$$

If the distributions of I and S are considered to be exponential with rates λ and μ respectively, then the system is known as M/M/1 queuing system and analytical solution E (W_i) is available as:

$$E(W_i) = 1/(\mu - \lambda)$$

The average number of customers in the system (waiting or being served) is $\rho / (1 - \rho)$. This may be verified by simulation of M/M/1 process.

16.2.2 Simulation of a Very Simple Inventory Problem: An Example

A newspaper vendor buys daily newspaper at Rs. 2.30 and sells them at Rs. 3.50 each. Newspapers are purchased at the beginning of the day (before the news paper boy knows what the demand of the papers will be), and any paper left unsold at the end of the day are thrown out.

Suppose the probability distribution of the random demand of newspapers is given as below. The news boy wishes to know how many papers he should buy each day in order to maximize his profit. The demand per day and their probabilities are given below:

Demand (D)	20	21	22	23	24	25	26	27	28	29	30
Probabilities	0.05	0.05	0.10	0.10	0.10	0.15	0.15	0.10	0.10	0.05	0.05

Suppose we begin with an inventory of arbitrary number, say, 25 papers purchased each day. We then evaluate the profit resulting from a policy of purchasing 25 papers at the beginning of each day. To do this, we fill out a table describing the actual sale and profits for each simulated day. The demand per day for 30 days can be simulated by the method of IPT and a simulation is given in the following table:

Day	1	2	3	4	5	6	7	8	9	10
Demand (D)	28	20	26	26	23	22	24	24	29	21
Profit (P)	30.0	12.5	30.0	30.0	23.0	19.5	26.5	26.5	30.0	16.0
Day	11	12	13	14	15	16	17	18	19	20
Demand (D)	24	24	28	28	26	27	26	26	24	21
Profit (P)	26.5	26.5	30.0	30.0	30.0	30.0	30.0	30.0	26.5	16.0
Day	21	22	23	24	25	26	27	28	29	30
Demand (D)	26	23	25	25	25	25	27	24	22	23
Profit (P)	30.0	23.0	30.0	30.0	30.0	30.0	30.0	26.5	19.5	23.0

$$\text{Sale} = D \quad \text{if } D \leq 25$$

$$= 25 \quad D > 25$$

$$\text{Profit} = \text{Sale} \times 3.50 - 25 - 2.30$$

Average Profit = Rs. 26.38

In the above example, we have given the profitability of ordering 25 papers. In this way one can calculate the profitability of ordering any other number and make a choice of ordering that number which gives the maximum profit.

E1) Simulate a M/M/1 process with $\lambda = 0.6$ and $\mu = 1.0$ and find out average waiting time W_i by taking N = 10.

E2) The following data give the arrival times and service times that each customer will require for the first 13 customers at a single server system:

Arrival Times:	12	31	63	95	99	154	198	221	304	346	411	455	537
Service Times:	40	32	55	48	18	50	47	18	28	54	40	72	12

a) Determine the waiting times of 13 customers.

b) Determine the average waiting time.

16.3 TEST FOR RANDOMNESS

In Section 13.3, we have seen that pseudo random numbers (PRN) generated are completely deterministic. In this section, we shall describe some statistical tests to see that how close they are to independent random numbers from U (0, 1) distribution. There are two quite different kind of tests as follows:

1. Theoretical tests
2. Empirical tests

Theoretical tests are based on parameters a, c and m of equation in a global manner without actually generating U_i's at all. This is outside the scope of this course and we shall not describe them here. We shall discuss empirical test in details in further sections.

In the following sections we shall consider two types of tests. The first type of tests i.e. Chi-square and Kolmogrov-Smirnov, are tests of goodness of fit**.** For given numbers they test whether they are from the assumed distribution e.g. Normal, Exponential, Poisson, etc. or not.

The second type of test i.e. Runs Up-and-Down test, is a test of independence. This tests whether the consecutive numbers which appear in a sequence are independent or not (i.e. consecutive numbers do not have any trend).

Some Empirical Tests

16.3.1 Chi-Square Goodness of Fit Test

Let $x_1, x_2, \ldots, x_n$ be a sample from a distribution F(x) which is unknown. We wish to test the null hypothesis

$$H_0: F(x) = F_0(x), \quad \text{for all } x$$

where, $F_0(x)$ is completely specified, against the alternative

$$H_1: F(x) \neq F_0(x), \quad \text{for some } x$$

Assume that n observations have been grouped into k mutually exclusive categories. Denote n_j and np_{j0} be the observed and expected number for the j^{th} category, j = 1, 2, …, k. Here p_{j0} is the probability of an observation lying in the j^{th} class and is calculated from $F_0(x)$.

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_{j0})^2}{np_{j0}},$$

where, $\sum_{j=1}^{k} n_j = n$

which should be small when H_0 is true and large when H_0 is false. The cut-off point between large and small is decided by distribution of statistic χ^2 which has approximately a Chi-square distribution on (k-1) degrees of freedom when H_0 is true and n is large.

Under H_0 we have

$$P(\chi^2 > \chi^2_\alpha) = \alpha$$

and then we find χ^2_α from table of Chi-square distribution.

We reject H_0 if $\chi^2 > \chi^2_\alpha$.The α should be taken = 0.10 or 0.05.

When testing for uniform distribution, we divide the [0,1] interval into k non-overlapping classes of equal length 1/k.

Now, under H_0: $np_{j0} = n/k$

The test statistic is

$$\chi^2 = k\sum_{j=1}^{k} \frac{(n_j - n/k)^2}{n},$$

Under null hypothesis χ^2 has a Chi-square distribution on (k-1) degrees of freedom. For good approximation of the distribution one should have n > 50,000 and $k = 4n^2/5$.

For a good approximation to Chi-square distribution the categories should be such that the expected frequencies in each class, np_{j0}, should be greater than 5. In case this is not so the consecutive classes should be merged to assure that the expected frequencies are greater than 5.

16.3.2 Kolmogrov-Smirnov Goodness of Fit Test

Let $x_1, x_2, ..., x_n$ denoete a random sample from unknown continuous distribution function F(x).The empirical distribution derived from sample, denoted by $F_n(x)$ is:

$$F_n(x) = (\text{Number of } x_i\text{'s} \leq x)/n$$

If null hypothesis

$$H_0: F(x) = F_0(x) \quad \text{for all } x$$

is correct then one would expect that deviation $| F_0(x) - F_n(x) |$ should be small for all values of x. The asymptotic distribution of D_n

$$D_n = \sup_x |F_0(x) - F_n(x)|$$

has been studied by Kolmogrov-Smirnov and the critical values, d_n, have been tabulated

$$P(D_n > d_{1-\alpha}) = \alpha$$

for various values of n.

In case $D_n > d_{1-\alpha}$ then we reject the null hypothesis at a significance level α. When $F_0(x)$ is uniform distribution U (0, 1) then D_n

$$D_n = \sup_x |x - F_n(x)|$$

and one can test the goodness of random numbers generated.

Example 1: Twenty uniform U (0, 1) numbers (x) generated by a generator are given as below:

x	0.867	0.778	0.741	0.480	0.441	0.095	0.096	0.442	0.273	0.140
Rank (r_i)	17	15	14	11	9	2	3	10	6	4
x	0.968	0.786	0.019	0.330	0.408	0.681	0.144	0.978	0.889	0.579
Rank (r_i)	19	16	1	7	8	13	5	20	18	12

Demonstrate Kolmogrov-Smirnov test on these numbers.

Solution: We have the calculations in following table:

x	0.867	0.778	0.741	0.480	0.441	0.095	0.096	0.442	0.273	0.140
Rank (r_i)	17	15	14	11	9	2	3	10	6	4
$F_n(x) = r/n$	0.850	0.750	0.700	0.550	0.450	0.100	0.150	0.500	0.300	0.200
$\|x - F_n(x)\|$	0.017	0.028	0.041	0.070	0.009	0.005	0.054	0.058	0.027	0.060
x	0.968	0.786	0.019	0.330	0.408	0.681	0.144	0.978	0.889	0.579
Rank (r_i)	19	16	1	7	8	13	5	20	18	12
$F_n(x) = r/n$	0.950	0.800	0.050	0.350	0.400	0.650	0.250	1.000	0.900	0.600
$\|x - F_n(x)\|$	0.018	0.014	0.031	0.020	0.008	0.031	0.106*	0.022	0.003	0.021

We also have $F_0(x) = x$, so

$$D_n = \sup_x |x - F_n(x)|$$

$$\approx$$

$$D_n = 0.106$$

$d_{0.95} = 0.29$ for n = 20 (from Kolmogrov-Smirnov Table)

$D_n = 0.106 < 0.29$ (the tabulated value)

Thus, null hypothesis is not rejected, i.e. there is no significant departure from uniform distribution.

16.3.3 Runs Up and Down Test (Test of Independence)

One important departure from randomness is the propensity for the occurrence of long monotonic sub-sequences. Run tests are designed to detect non random behaviour of this type. A run is a monotonic subsequence for assessing the randomness 'runs up', 'runs down' and 'runs up and down' may be considered. If it is too large or small than the number of runs expected in a sequence of independent numbers, then it indicates its departure from independence. Consider an example of a sequence of nine numbers:

1,	5,	4,	1,	3,	1,	3,	4,	7
+	–	–	+	–	+	+	+	+

+ sign is given when the next number is larger than the previous one and

– sign, when the next number is smaller than the previous one. This sequence contains three runs up (+) and two runs down (-) thus making a total, T, of five runs (T = 5).

Under independence, this statistic T is approximately normally distributed with mean and variance

$$E(T) = (2n-1)/3,$$

$$V(T) = (16n - 29)/90$$

for n greater than 25. Test statistic Z:

$$Z = \frac{T - \frac{(2n-1)}{3}}{\left[\frac{(16n-29)}{90}\right]^{1/2}}$$

has a standard normal distribution under null hypothesis and test can be applied.

One potential disadvantage of empirical test is that they are only local in nature i.e. only that segment of a cycle (of LCG for example) which was actually used, is tested for randomness, therefore we cannot say how the generator might perform in other segments of the cycle. On the other hand, this local nature of empirical tests can be an advantage, since it might allow us to examine the actual random numbers that will be used in simulation.

Example 2: A random number generator produces the following U(0,1) random numbers:

0.34	0.50	0.04	0.75	0.76	0.61	0.66	0.32	0.48	0.94
+	-	+	+	-	+	-	+	+	+
0.19	0.18	0.49	0.39	0.66	0.48	0.21	0.07	0.88	0.80
-	-	+	-	+	-	-	-	+	-
0.31	0.06	0.88	0.27	0.31	0.64	0.86	0.93	0.57	0.82
-	-	+	-	+	+	+	+	-	+
0.76	0.68	0.61	0.49	0.13	0.92	0.03	0.21	0.22	0.17
-	-	-	-	-	+	-	+	+	-

Apply Chi-square goodness of fit test to test that random numbers come from uniform distribution and successive numbers are independent.

Solution: We construct the following frequency table**:**

Class	Frequency (n_j)
0.0 - 0.2	8
0.2 - 0.4	9
0.4 - 0.6	6
0.6 - 0.8	9
0.8 - 1.0	8

Chi-square Goodness of Fit Test

Here one is interested in testing whether these numbers can be considered to be coming from a U(0,1) population.

Under uniform distribution $p_1 = p_2 = p_3 = p_4 = p_5 = 1/5$, $n = 40$,

$$\chi^2 = \sum_{j=1}^{5} \frac{(n_j - 8)^2}{8}$$

$$= (0^2 + 1^2 + 2^2 + 1^2 + 0^2) / 8$$

$$= 6 / 8 = 0.75$$

Therefore, $0.75 < \chi^2_{4,\,0.05} = 9.488$

Hence, the null hypothesis is accepted i.e. the hypothesis of Uniform distribution is not rejected.

Run Test

Here one is interested in testing whether the consecutive numbers can be considered as independent.

Total number of runs (up and down) T = 24 and n = 40.

$E(T) = 79/3 = 26.333,$

$V(T) = 6.789,$

$$Z = \frac{(24 - 26.333)}{\sqrt{6.789}}$$

$$= \frac{(24 - 26.333)}{2.605} = -0.89$$

We have $\alpha/2 = 0.025$.

From table $Z_{\alpha/2} = Z_{0.025} = 1.96$

$$|Z| = 0.89 < Z_{0.025} = 1.96$$

Hence we do not reject the null hypothesis of randomness. Thus, we conclude that successive observations are independent.

E3) Following U (0, 1) were generated by a generator. Apply Chi-square goodness of fit test to test the fit of the distribution.

0.151	0.669	0.053	0.475	0.290	0.235	0.966	0.004	0.100
0.919	0.497	0.729	0.912	0.822	0.591	0.717	0.964	0.439
0.319	0.205	0.349	0.898	0.819	0.370	0.272	0.751	0.543
0.409	0.813	0.345	0.364	0.347	0.179	0.313	0.394	0.872
0.883	0.332	0.268	0.474					

E4) The following table gives the grouped data for number of items demanded per day. They were generated by Poisson distribution algorithm with mean $\lambda = 6$. By using Chi-square goodness of fit statistic test the hypothesis that the generated data fit the Poisson distribution with $\lambda = 6$.

Demand (X)	:≤3	4	5	6	7	8-9	≥10
Frequency (n_j):	12	10	12	18	10	20	5

E5) The following table gives the frequency distribution of 100 variables generated from N (0, 1) distribution. Using Chi- square goodness of fit test statistics test whether fit is satisfactory.

Class Interval	Frequency
≤ (-2.5)	02
(-2.5) - (-1.5)	04
(-1.5) - (-1.0)	08
(-1.0) - (-0.5)	18
(-0.5) - 0	19
0 - (0.5)	12
(0.5) - (1.0)	14
(1.0) - (1.5)	14
(1.5) - (2.0)	05
(2.0) - (2.5)	02
(2.5) - (3.0)	02

E6) Times between successive crashes of a computer system were generated for a 6-month period and are given in increasing order as follows (time in hours):

1 10 20 30 40 52 63 70 80 90 100 102 130 140 190 210 266 310 530 590 640 1340

The parameter $\alpha = 0.00435$, Mean $= 1/\alpha = 230$ hrs.

Use Kolmogrov-Smirnov test to examine the goodness of fit of Exponential distribution.

16.4 ADVANTAGES AND DISADVANTAGES OF SIMULATION

By using simulation techniques it is possible to study and experiment with the complex internal interactions of a given system whether it is a firm, an industry or economy.

It is possible to study the effect of certain changes which may lead to improvement of the system.

Simulation can be used to experiment with new situation about which we have little or no information. This is a very powerful tool for decision making. In some situations it may be used to verify analytical solutions and obtain solutions of problems which are difficult to solve analytically.

Simulation is a very useful technique for those situations where analytical techniques are inadequate. It has a number of disadvantages too. It is an imprecise technique which is subject to random variations rather than exact results. It gives results for the parameter values actually used in simulation. It may not be advisable to generalise those results which have not been covered in the simulation. Usefulness of the simulation depends on the adequacy of the model it represents. In case it is not correct representation of the system then conclusions drawn may not be valid for the system under study.

16.5 SUMMARY

In this unit, we have discussed:

1. Areas where simulation can be applied to solve certain problems which are very difficult to solve otherwise;
2. The queuing and inventory problems in particular which have many applications of simulation;
3. The advantages and disadvantages of simulation; and
4. Some empirical tests of randomness of a sequence of random numbers generated.

16.6 SOLUTIONS /ANSWERS

E1) Ten U (0, 1) variates given in the following table are:

U	0.34	0.5	0.04	0.75	0.76	0.61	0.66	0.32	0.48	0.94
$I = (-\log U)/\lambda$	1.80	1.15	5.36	0.48	0.46	0.82	0.69	1.90	1.22	0.10
U	0.19	0.18	0.49	0.39	0.66	0.48	0.21	0.07	0.88	0.87
$S= (-\log U)/\mu$	1.66	1.71	0.71	0.94	0.41	0.73	1.56	2.66	0.13	0.14

Using the formula

$$W_n = W_{n-1} - I_n + S_n \quad \text{if } W_{n-1} > I_n,\ n = 1, 2, \ldots$$
$$= S_n \quad \text{if } W_{n-1} < I_n$$
$$W_1 = S_1$$

gives,

W: 1.66 2.22 0.71 1.17 1.12 1.03 1.90
2.66 1.57 1.61

Average, $\bar{W} = 1.565$

Theoretical value = $1/(\mu - \lambda) = 1/0.4 = 2.5$

E2) We have $I_i = t_i - t_{i-1}$

$$W_n = W_{n-1} - I_n + S_n \quad \text{if } W_{n-1} > I_n,\ n = 1, 2, \ldots$$
$$= S_n \quad \text{if } W_{n-1} < I_n$$
$$W_1 = S_1$$

I_i	12	19	32	32	4	55	44	23	83	42	65	44	82
S_i	40	32	55	48	18	50	47	18	28	54	40	72	12
W_i	40	53	76	92	106	101	104	99	44	56	40	72	12

Therefore, $\bar{W} = 68.85$

E3) Chi-square goodness of fit test:

Class Interval	Class Frequency (n_j)	np_{j0}
0-0.2	5	8
0.2-0.4	14	8
0.4-0.6	7	8
0.6-0.8	4	8
0.8-1.0	10	8

Expected frequency = 40/5 = 8 = np_{j0}

Then $$\chi^2 = \sum_{j=1}^{5} \frac{(n_j - np_{j0})^2}{np_{j0}} = \sum_{j=1}^{5} \frac{(n_j - 8)^2}{8}$$

$$= (3^2 + 6^2 + 1^2 + 4^2 + 2^2)/8$$

$$= 8.25 < \chi^2_{4,0.05} = 9.488$$

Hence, the null hypothesis is not rejected, i.e. there is no significant departure from the Uniform distribution.

E4) Following table gives the probability p_j for each class (obtained from Poisson table) with $\lambda = 6$:

p_{j0}	:	0.1512	0.1339	0.1606	0.1606	0.1377	0.1721	0.0839
np_{jo}	:	13.15	11.65	13.97	13.97	11.98	14.97	7.30
$n_j - np_{j0}$	:	-1.15	-1.65	-1.97	4.03	-1.98	5.03	-2.30

Therefore, $$\chi^2 = \sum_{j=1}^{5} \frac{(n_j - np_{j0})^2}{np_{j0}} = 4.517$$

Tabulated value of Chi-Square on 6 degrees of freedom for $\alpha = 0.05$, is $\chi^2_{6,\,0.05} = 12.6$.

Thus calculated $\chi^2 < \chi^2_{6,\,0.05}$. Hence we do not reject the null hypothesis i.e. the data generated give satisfactory fit.

E5) Seeing the Normal tables we have obtained probability of each class given as p_{j0}

Class Interval	p_{j0}	np_{j0}		Observed Frequency (n_j)		$n_j - np_{j0}$
≤ 25	0.0062	00.62	6.68	02	06	0.68
(-2.5, -1.5)	0.0606	06.06		04		
(-1.5, -1.0)	0.0918	09.18		08		1.18
(-1.0, -0.5)	0.1499	14.99		18		-3.01
(-0.5, 0.0)	0.1915	19.15		19		0.15
(0.0, 0.5)	0.1915	19.15		12		7.15
(0.5, 1.0)	0.1499	14.99		14		0.99
(1.0, 1.5)	0.0918	9.18		14		-4.82
(1.5, 2.0)	0.0441	4.41	7.23	05	09	-1.77
(2.0, 2.5)	0.0165	1.65		02		
(2.5, 3.0)	0.0117	1.17		02		

We have merged two classes in the top of the table and three in the bottom of the table so as to get better approximation to the Chi-square distribution. The calculated value of the test statistic is

$$\chi^2 = \sum_{j=1}^{8} \frac{(n_j - np_{j0})^2}{np_{j0}} = 6.52$$

Tabulated value of Chi-Square on 7 degrees of freedom for $\alpha = 0.05$ is

$$\chi^2_{7,\,0.05} = 14.1$$

Thus $\chi^2 < 14.1$. Hence, we do not reject the null hypothesis and conclude that the fit is satisfactory.

E6) For the Exponential distribution

$$P(X \le x) = F_0(x) = 1 - e^{-\alpha x} = 1 - e^{-0.00435x}$$

The following table gives the required calculations $F_n(x) = x/22$

S.No.	1	2	3	4	5	6	7	8	9	10	11
x	1	10	20	30	40	52	63	70	80	90	100
$F_n(x)$	0.045	0.091	0.136	0.182	0.227	0.273	0.318	0.364	0.409	0.454	0.500
$F_0(x)$	0.009	0.051	0.032	0.122	0.160	0.202	0.240	0.262	0.294	0.324	0.353
$\lvert F_n(x)-F_0(x)\rvert$	0.036	0.040	0.104	0.060	0.067	0.071	0.078	0.102	0.115	0.130	0.147
S.No.	12	13	14	15	16	17	18	19	20	21	22
x	102	130	140	190	210	266	310	530	590	640	1340
$F_n(x)$	0.545	0.591	0.636	0.682	0.727	0.773	0.818	0.864	0.909	0.954	1.000
$F_0(x)$	0.358	0.432	0.456	0.562	0.599	0.686	0.740	0.900	0.923	0.938	0.997
$\lvert F_n(x)-F_0(x)\rvert$	0.187*	0.159	0.180	0.120	0.128	0.087	0.078	0.036	0.014	0.016	0.003

$$D_n = \sup_x |F_0(x) - F_n(x)| = 0.187,$$

$\approx$

and $d_{22,0.05} = 0.28$ (from table)

D_n observed is smaller than tabulated value of 0.28 at $\alpha = 0.05$. Hence, we do not reject the null hypothesis. The conclusion is that it gives a satisfactory fit.

TABLE 1: Random Number Table

03339	19233	50911	14209	39594	68368	97742	36252	27671	55091
97971	19968	31709	40197	16313	80020	01588	21654	50328	04577
16779	47712	33846	84716	49870	59670	46946	71716	50623	38681
12675	95993	08790	13241	71260	16558	83316	68482	10294	45137
55804	72742	16237	72550	10570	31470	92612	94917	48822	79794
16835	56263	53062	71543	67632	30337	28739	17582	40924	32434
84544	14327	07580	48813	30161	10746	96470	60680	63507	14435
63230	41243	90765	08867	08033	05038	10908	00633	21740	55450
33564	93563	10770	10595	71323	84243	09402	62877	49762	56151
57461	55618	40570	72906	30794	49144	65239	21788	38288	29180
91645	42451	83776	99246	45548	02457	74804	49536	89815	74285
78305	63797	26995	23146	56071	97081	22376	09819	56855	97424
97888	55122	65545	02904	40042	70653	24483	31258	96475	77668
67286	09001	09718	67231	54033	24185	52097	78713	95910	84400
53610	59459	89945	72102	66595	02198	26968	88467	46939	52318
52965	76189	68892	64541	02225	09603	59304	38179	75920	80486
25336	39735	25594	50557	96257	59700	27715	42432	27652	88151
73078	44371	77616	49296	55882	71507	30168	31876	28283	53424
31797	52244	38354	47800	48454	43304	14256	74281	82279	28882
47772	22798	36910	39986	34033	39868	24009	97123	59151	27583
54153	70832	37575	31898	39212	63993	05419	77565	73150	98537
93745	99871	37129	55032	94444	17884	27082	23502	06136	89476
81686	51330	58828	74199	87214	13727	80539	95037	73536	16862
79788	02193	33250	05865	53018	62394	56997	41534	01953	13763
92112	61235	68760	61201	02189	09424	24156	10368	26257	89107
87542	28171	45150	75523	66790	63963	13903	68498	02891	25219
37535	48342	48943	07719	20407	33748	93650	39356	01011	22099
95957	96668	69380	49091	90182	13205	71802	35482	27973	46814
34642	85350	53361	63940	79546	89956	96836	91313	80712	73572
50413	31008	09231	46516	61672	79954	01291	72278	55658	
84893									
53312	73768	59931	55182	43761	59424	79775	17772	41552	45236
16302	64092	76045	28958	21182	30050	96256	85737	86962	27067
96357	98654	01909	58799	87374	53184	87233	55275	59572	56476
38529	89095	89538	15600	33687	86353	61917	63876	52367	79032
45939	05014	06099	76041	57638	55342	41269	96173	94872	35605
02300	23739	68485	98567	77035	91533	62500	31548	09511	80252
59750	14131	24973	05962	83215	25950	43867	75213	21500	17758
21285	53607	82657	22053	88931	84439	94747	77982	61932	21928
93703	60164	19090	63030	88931	84439	94747	77982	61932	21928
15576	76654	19775	77518	43259	82790	08193	63007	68824	75315
12752	33321	69796	03625	37328	75200	77262	99004	96705	15540
89038	53455	93322	25069	88186	45026	31020	52540	10838	72490
62411	56968	08379	40159	27419	12024	99694	68668	73039	87682
45853	68103	38927	77105	65241	70387	01634	59665	30512	66161
84558	24272	84355	00116	68344	92805	52618	51584	75901	53021
45272	58388	69131	61075	80192	45959	76992	19210	27126	45525
68015	99001	11832	39832	80462	70468	89929	55695	77524	20675
13263	92240	89559	66545	06433	38634	36645	22350	81169	97417
66309	31446	97705	46996	69059	33771	95004	89037	38054	80853
56348	05291	38713	82303	26293	61319	45285	72784	50043	44438

TABLE 1 (Continued)

93108	77033	68325	10160	38667	62441	87023	94372	06164	30700
28271	08589	83279	48838	60935	70541	53814	95588	05832	80235
21841	35545	11148	25255	50283	94037	57463	92925	12042	91414
09210	20779	02994	02258	86978	85092	54052	18354	20914	28460
90552	71129	03621	20517	16908	06668	29916	51537	93658	29525
01130	06995	20258	10351	99248	51660	38861	49668	74742	47181
22604	56719	21784	68788	38358	59827	19270	99287	81193	43366
06690	01800	34272	65479	94891	14537	91358	21587	95765	72605
59809	69982	71809	64984	48709	43991	24987	69246	86400	29559
56475	02726	58511	95405	70293	84971	06676	44075	32338	31980
02730	34870	83209	03138	07715	31557	55242	61308	26507	06186
74482	33990	13509	92588	10462	76546	46097	01825	20153	36271
19793	22487	94238	81054	95488	23617	15539	94335	73822	93481
19020	27856	60526	24144	98021	60564	46373	86928	52135	74919
69565	60635	65709	77887	42766	86698	14004	94577	27936	47220
69274	23208	61035	84263	15034	28717	76146	22021	23779	98562
83658	14204	09445	40430	54072	82164	68977	95583	11765	81072
14980	74158	78216	38985	60838	82806	49777	85321	90463	11813
63172	28010	29405	91554	75195	51183	65805	87525	35952	83204
71167	37984	52737	06869	38122	95322	41356	19391	96787	64410
78530	56410	19195	34434	83712	20758	83454	22756	83959	96347
98324	03774	07573	67864	06497	20758	83454	22756	83959	96347
55793	30055	08373	32652	02654	75980	02095	87545	88815	80086
05674	34471	61967	91266	38814	44728	32455	17057	08339	93997
15643	22245	07592	22078	73628	60902	41561	54608	41023	98345
66750	19609	70358	03622	64898	82220	69304	46235	97332	64539
42320	74314	50222	82339	51564	42885	50482	98501	00245	88990
73752	73818	15470	04914	24936	65514	56633	72030	30856	85183
97546	02188	46373	21486	28221	08155	23486	66134	88799	49496
32569	52162	38444	42004	78011	16909	94194	79732	47114	23919
36048	93973	82596	28739	86985	58144	65007	08786	14826	04896
40455	36702	38965	56042	80023	28169	04174	65533	52718	55255
33597	47071	55618	51796	71027	46690	08002	45066	02870	60012
22828	96380	35883	15910	17211	42358	14056	55438	98148	35384
00631	95925	19324	31497	88118	06283	84596	72091	53987	01477
75722	36478	07634	63114	27164	15467	03983	09141	60562	65725
80577	01771	61510	17099	28731	41426	18853	41523	14914	76661
10524	20900	65463	83680	05005	11611	64426	59065	06758	02892
93185	69446	75253	51915	97839	75427	90685	60352	96288	34248
81867	97119	93446	20862	46591	97677	42704	13718	44975	67145
64649	07689	16711	12169	15238	74106	60655	56289	74166	78561
55768	09210	52439	33355	57884	36791	00853	49969	74814	09270
38080	49460	48137	61589	42742	92035	21766	19435	92579	27683
22360	16332	05343	34613	24013	98831	17157	44089	07366	66196
40521	09057	00239	51284	71556	22605	41293	54854	39736	05113
19292	40078	06838	05509	68581	39400	85615	52314	83202	40313
64138	27983	84048	42635	58658	62243	82572	45211	37060	15017

TABLE 2: The χ^2 Table

The first column identifies the specific χ^2 distribution according to its number of degrees of freedom. Other columns give the proportion of the area under the entire curve which falls above the tabled value of χ^2.

Area in the Upper Tail

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.00039	0.00016	0.00098	0.0039	0.016	2.71	3.84	5.02	6.63	7.88
2	0.010	0.020	0.051	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.072	0.11	0.22	0 .35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.05	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	107.56	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

TABLE 3: Critical values of D in the Kolmogorov-Smimov One Sample Test

Sample size	Level of significance of d = Sup $\lvert F_0(x) - F_n(x) \rvert$				
(n)	.20	.15	.10	.05	.01
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.383	.432	.514
10	.322	.342	.368	.410	.490
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.392
17	.250	.266	.286	.318	.381
18	.244	.259	.278	.309	.371
19	.237	.252	.272	.301	.363
20	.231	.246	.264	.294	.356
25	.21	.22	.24	.27	.32
30	.19	.20	.22	.24	.29
35	.18	.19	.21	.23	.27
Over 35	1.07	1.14	1.22	1.36	1.63

UNIT 15 SIMULATION TECHNIQUES

Structure

15.1 Introduction
Objectives
15.2 Steps in Simulation
15.3 Monte-Carlo Method of Simulation
Monte-Carlo Integration
Simulation in Statistical Inference
Estimation of a Distribution of a Parameter by Monte-Carlo Simulation
Problem of Discrete Event Simulation
Monte-Carlo Simulation in Business Applications
15.4 Summary
15.5 Solutions / Answers

15.1 INTRODUCTION

Dictionary meaning of simulation is to pretend, to act like or mimic, etc. Some of the examples of simulation are: model of a car, aeroplane in a wind tunnel may act like in a real flight, etc. More precisely a simulation of a system is the operation of a model (or simulator) which is the representation of the system. The model is amenable to manipulation, which would be impossible, too expensive or impractical to perform on the entity it portrays. The operation of the model can be studied and from it, properties concerning the behaviour of the actual system (or its subsystems) can be inferred. Some examples are forest management, epidemics, traffic congestion, effect of ageing in a population, etc.

The steps involved in simulation are described in Section 15.2. The Monte-Carlo method of simulation is explained in Section 15.3. In the same section we have discussed Monte-Carlo integration, simulation in statistical inference, estimation of a distribution of a parameter by Monte-Carlo simulation, problem of discrete event simulation and Monte-Carlo simulation in business applications

Objectives

After studying this unit, you would be able to

- define the simulation;
- describe different steps in setting up simulation for solving a problem;
- describe the Monte-Carlo simulation;
- use Monte-Carlo simulation for solving some deterministic and stochastic problems; and
- describe some methods for making inference by simulation.

15.2 STEPS IN SIMULATION

The following are the principle steps in simulation:

1. **Formulation of the Problem and Plan the Study**

   It needs clear statement of the problem and its objectives. If there are some alternatives then they should be documented. Criteria for comparing efficiencies of alternatives should be given. It should also plan for cost and time required for the study.

2. **Collect Data and Define a Model**

   One has to collect data from the system of interest and also obtain the estimate of parameters of the model. Choice of the model is very important. As a rule one should start with a simple model and then make it more complex as the need be.

3. **Validation of the Model**

   It is very important that the model chosen should be true representative of the system of interest. This needs verification at every stage of the operation. The adequacy of the theoretical probability distribution fitted to the observed data should be tested using goodness of fit tests.

4. **Construction of a Computer Program**

   These days most of the simulation work is done on computers. For this, program is to be written and errors removed.

5. **Make Pilot Runs of the Computer Program and Check the Output for Validity**

   It is important to check the program in the pilot runs by running it for some cases where correct results are known.

6. **Make Production Runs**

7. **Analyse Output Data**

8. **Document and Implement Results**

15.3 MONTE-CARLO METHOD OF SIMULATION

One may define simulation as a numerical technique for conducting experiments on a digital computer, which involves certain types of mathematical and logical models that describe the behaviour of the system of interest. Here we are concerned with stochastic simulation, which is also called Monte-carlo simulation. This involves sampling stochastic variables from some probability distribution. Some problems which are stochastic in nature, such as consumer demand, production, population size, number of persons waiting in a queue, etc. are some examples of this type.

Some completely deterministic problems are very difficult to solve analytically. However, by simulating a stochastic process, whose moments, density function or cumulative distribution function satisfy the solution requirements, give approximate solution of the problem. Some examples of this nature are solution of certain integral and differential equations. In this

section, we shall outline some important simulation techniques in different fields.

15.3.1 Monte-Carlo Integration

Suppose we wish to evaluate a deterministic integral

$$\theta = \int_0^1 g(x)\,dx$$

If this integral exists then it is merely E[g(x)] where x is U (0,1) variate. Hence θ can be estimated as $\hat{\theta}$ given by

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

where $x_1, x_2, \ldots, x_n$ is a random sample of size n from U(0,1) population. $\hat{\theta}$ is an unbiased estimate of θ. As n increases, $\hat{\theta}$ approaches θ, and thus accuracy of $\hat{\theta}$ can be increased by increasing n. This is an example of solving a deterministic problem by Monte-carlo simulation.

15.3.2 Simulation in Statistical Inference

Suppose we have a sample $x_1, x_2, \ldots, x_n$ from a probability distribution function f (x, θ) where θ is not known. We wish to test the hypothesis

$$H_0: \theta = \theta_0$$

against the alternative hypothesis

$$H_1: \theta > \theta_0$$

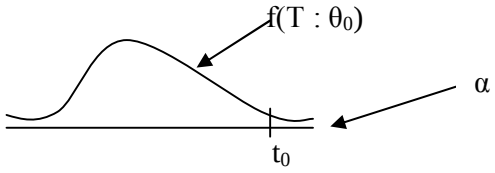
Suppose, we have a test statistic $T = g(x_1, x_2, ..., x_n)$, where g is a known function of x_i's. Suppose large values of T indicate the departure from the null hypothesis. Let $t = g(x_1, x_2, ..., x_n)$ be the observed statistic from the sample.

If the theoretical distribution of T, f (T, θ_0) under $H_0: \theta = \theta_0$, is known and if

$$P(T > t_0) = \alpha,$$

then if $T > t_0$, we reject H_0. Accept H_0 otherwise.

However, if f (T, θ_0) is not known then one can always simulate the distribution of T by taking many samples from f(x, θ_0) and calculating t for each sample and thus obtaining an empirical distribution of T. Then one can estimate critical point t_0 and test the hypothesis. Two sided hypothesis can also be tested once the empirical distribution is available.



Example 1: A random sample of size ten from a population is given as follows:

8.94, 6.31, 5.00, 3.94, 5.19, 3.80, 6.11, 5.65, 3.76, 5.03

Assuming that population can be considered as Normal with standard deviation 1, test the hypothesis H_0 that the mean of the population is 5 against the alternative hypothesis H_1 that it is greater than 5, by using statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad \text{where} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Solution: In this case the exact distribution of statistic t is known, but we shall demonstrate the test by using empirical distribution of t. For this we have generated 100 random samples of size 10 from normal population with mean $\mu_0 = 5$ and $\sigma = 1$ by the methods described in Unit 14 and calculated t for all 100 samples. We then made a frequency table. Value of t calculated from the given sample is

t	frequency
-0.2 -0.15	7
-0.15 -.0.1	5
-0.1 -.0.05	15
-0.05 -0.0	25
0 -0.05	16
0.05 -0.10	16
0.10 -0.15	9
0.15 -0.2	5

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$= \frac{5.373 - 5}{1/\sqrt{10}}$$

$$= 1.179$$

From the frequency table, we observed that there are 14 t's greater than observed1.179. Hence this is a large prbability i.e. p=0.14(=14/100) thus we do not reject the null hypothesis of mean of the population being 5.

Hence, we used the theoretical distribution of t, which is N (0, 1) and from normal table p = 0.12 = P (Z > 1.179). In both cases, we accept the null hypothesis. The values of p differ because empirical distribution is not very accurate due to small number of samples (100) involved in generating it. Usually more than 1000 samples are required for estimating empirical distribution.

15.3.3 Estimation of a Distribution of a Parameter by Monte-Carlo Simulation

Suppose the purpose is to find the distribution of some of the parameters of the distribution of a random variable. The random variable, which we shall call output variable, is a known function of other random variables which have known distributions. To estimate the distribution of the output variable we draw a value for each of the input variables from their distributions and calculate the resulting output variable. Such sampling is then repeated many times and this yields an estimate of the distribution.

Example 2: Using Monte-Carlo method, estimate p where

$$p = P[g(x) < a]$$

where, $g(x) = \min(x_1, x_2)$

Here x_1 and x_2 have independent normal distributions,

$$x_1 \sim N(100, 400) \quad ; \quad x_2 \sim N(90, 100)$$

Solution: The problem of the estimation of p arises if we have a product consisting of two parts. The distribution of lives of part 1 and part 2, say x_1 and x_2 respectively, are given above. The product breaks down as soon as one of the two parts fails. We want to know the probability that the life of the product is smaller than some given value, say, a. Therefore, we want to know

p given above. For Monte-Carlo estimation of p, it is convenient to introduce a variable y defined as:

$$y = 1 \quad \text{if } g(x) < a$$
$$= 0 \quad \text{if } g(x) \geq a$$

The expected value of y, E(y) is given by

$$E(y) = 1.\ P\ (g(x) < a) + 0.\ P\ (g(x) \geq a)$$
$$= P\ (g(x) < a) = p$$

Now, we generate two independent random variables $x_1 \sim N\ (100, 400)$ and $x_2 \sim N\ (90, 100)$ as described in Unit 14, using independent U (0, 1) variables and find y. We repeat this a large number of times say, N, and then find y_i (i = 1, 2, ..., N) in each case . We estimate p as

$$\hat{p} = \bar{y} = \sum_{i=1}^{N} y_i / N$$

It is not necessary that x_1 and x_2 have normal distributions. One may consider any distribution and then estimate p as given above.

15.3.4 Problem of Discrete-Event Simulation

A very important example of discrete event simulation involves a queue (a waiting line). There are many practical applications of queuing problems. For example, banks, doctor's offices, supermarkets, car washers, airports and gasoline stations all involve situations in which customers may have to wait in line for service. A similar problem applies to orders entering a job shop which must wait for their turn to be processed, machine waiting to be repaired, jobs entering a computer system and so on. There are many situations in the queuing problem. Simplest situation is one in which there is only one single waiting line followed by a single server. Then there are cases where multiple servers (and multiple waiting lines) in series. Then there is a problem of multiple servers in parallel with a common waiting line. We shall discuss a simple example of this in Sub-section 16.3.1 with single line and single server.

15.3.5 Monte-Carlo Simulation in Business Applications

We shall consider a simple problem of simulation in assessing financial risk. Risk analysis is one of the most important and widely used applications of discrete event simulation. The objective is to assess the desirability of a proposed investment, based upon some financial decision criterion such as present worth. Application of the method results in a cumulative distribution being generated for the decision criterion. Hence, we can obtain not only the expected value of the decision criterion but also probabilities of obtaining much higher or lower values.

Suppose we are considering an initial outlay that will generate a series of n yearly cash flows (i.e. inflows and outflows). The present value of each cash flow can be written as

$$(PW)_j = \frac{(YCF)_j}{(1+i)^j},$$

where, $(YCF)_j$ represents the annual cash flow for the j^{th} year in the future and i is a specified annual interest rate, expressed as a decimal value. Hence the present worth of the entire proposed investment can be expressed as

$$PW = \left\{\sum_{j=1}^{n}(PW)_j\right\} - I$$

$$= \sum_{j=1}^{n}\frac{(YCF)_j}{(1+i)^j} - I$$

where I represents the initial cash outlay.

Each of the yearly cash flows is generally comprises of several components, such as yearly sales volume, production cost, taxes, etc. These items are normally represented in terms of appropriate distribution functions. Thus, the computational procedure involves generating a random value for each cash flow component, resulting in a randomly generated value of $(PW)_j$. All the yearly cash flows are evaluated in this manner. These values are then used in above equation resulting in a single value of PW. We then repeat the entire procedure N times, which allows us to obtain, a distribution for PW. We may calculate the average PW and other quantiles of the distribution.

Example 3: Suppose an industry has started a factory with an initial capital of Rs. 50 million. Suppose $(YCF)_j$ is distributed normally with mean 10 million and variance of 1 million^2. For one simulation (N =1) obtain PW for n =10, taking annual interest of 10 percent (i = 0.10).

Solution: Suppose ten normal variables N (0, 1) are given as the following:

N(0,1)	-0.179	0.421	0.210	-1.598	1.717	0.308	-0.421	-0.776	0.640	-0.319
N(10,1) $(YCF)_j$ in million rupees	9.821	10.210	10.210	8.402	11.717	9.692	9.579	9.224	10.640	9.681
$(1+i)^j$	1.100	1.210	1.331	1.464	1.610	1.771	1.949	2.143	2.358	2.594
$(PW)_j$	8.928	8.438	7.671	5.739	7.278	5.473	4.915	4.304	4.512	3.732

$$PW = \left\{\sum_{j=1}^{n}(PW)_j\right\} - I$$

$$PW = \sum_{j=1}^{10}(PW)_j - 50.000$$

$$= 60.990 - 50.000$$

$$= 10.990$$

Hence, for this simulation the worth is Rs 10.990 million. Such simulations are repeated a large number of times N say, more than 200, and distribution of PW is estimated. Every time new set of random normal variables are generated.

E1) By generating 10 uniform random variate U (0, 1) estimate the integral

$$\theta = \frac{1}{\sqrt{2\pi}} \int_{-1}^{2} e^{-x^2/2}\, dx$$

Recognizing this function as probability density function of N (0, 1), compare the value of $\hat{\theta}$ with θ.

E2) Suppose it is known that height of adult male population can be approximated by a normal distribution with mean of 1.62 meters and variance of 0.14 meters^2.

Using random numbers generated in E1), simulate the height of ten persons from the N (1.62, 0.14). Estimate the mean variance and range from this sample.

15.4 SUMMARY

In this Unit, we have discussed:

1. The nature and problems that can be solved by simulation;
2. Various steps involved in solving problems by simulation;
3. How approximate solution of purely deterministic problems can be obtained;
4. Some methods for inference based on simulation; and
5. Some examples of simulations.

15.5 SOLUTIONS/ANSWERS

E1) We have given the function

$$g(x) = (2\pi)^{-1/2} e^{-x^2/2}$$

If $x \sim U(-1, 2)$ then

$$E[g(x)] = \left(\frac{1}{2+1}\right) \frac{1}{\sqrt{2\pi}} \int_{-1}^{2} e^{-x^2/2}\, dx$$

Then $\theta = 3\, E[g(x)]$.

If we generate ten random variables from U (-1, 2) then an estimator $\hat{\theta}$ is given by

$$\hat{\theta} = 3 \sum_{i=1}^{10} \frac{g(x_i)}{10}$$

Ten simulated random U (0, 1) for a LCG were obtained as:

0.222, 0.198, 0.168, 0.784, 0.033, 0.932, 0.788, 0.237, 0.154, 0.587

$x \sim U(-1, 2)$ are obtained by $x = 3u - 1$,

and are given by

$-0.334, -0.406, -0.496, 1.352, -0.901, 1.796, 1.364, -0.289, -0.538, 0.761$

We have $\theta = 3 E[g(x)]$

$E[g(x)]$ is estimated as

$$E[g(x)] = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{10} \frac{e^{-x^2/2}}{10}$$

$$= 0.2786$$

Therefore, $\hat{\theta} = 3 \times 0.2786 = 0.8360$

So, θ from Normal distribution is given by

$$\theta = 0.8185 \text{ (from normal table)}$$

Hence, $\hat{\theta}$ is not very good estimate of θ. Perhaps increase in sample size, which is ten here, will give a better estimate.

E2) We have N (1.62, 0.14) normal variables generated by Box-Muller transformation on U (0, 1) variables, given in E1) are

1.71, 1.80, 1.58, 1.49, 1.78, 1.35, 1.63, 1.70, 1.39, 1.58

Then,

Mean = 1.60m,

Standard Deviations = 0.154 m

Range = 0.45 m

TABLE 1: Random Number Table

03339	19233	50911	14209	39594	68368	97742	36252	27671	55091
97971	19968	31709	40197	16313	80020	01588	21654	50328	04577
16779	47712	33846	84716	49870	59670	46946	71716	50623	38681
12675	95993	08790	13241	71260	16558	83316	68482	10294	45137
55804	72742	16237	72550	10570	31470	92612	94917	48822	79794
16835	56263	53062	71543	67632	30337	28739	17582	40924	32434
84544	14327	07580	48813	30161	10746	96470	60680	63507	14435
63230	41243	90765	08867	08033	05038	10908	00633	21740	55450
33564	93563	10770	10595	71323	84243	09402	62877	49762	56151
57461	55618	40570	72906	30794	49144	65239	21788	38288	29180
91645	42451	83776	99246	45548	02457	74804	49536	89815	74285
78305	63797	26995	23146	56071	97081	22376	09819	56855	97424
97888	55122	65545	02904	40042	70653	24483	31258	96475	77668
67286	09001	09718	67231	54033	24185	52097	78713	95910	84400
53610	59459	89945	72102	66595	02198	26968	88467	46939	52318
52965	76189	68892	64541	02225	09603	59304	38179	75920	80486
25336	39735	25594	50557	96257	59700	27715	42432	27652	88151
73078	44371	77616	49296	55882	71507	30168	31876	28283	53424
31797	52244	38354	47800	48454	43304	14256	74281	82279	28882
47772	22798	36910	39986	34033	39868	24009	97123	59151	27583
54153	70832	37575	31898	39212	63993	05419	77565	73150	98537
93745	99871	37129	55032	94444	17884	27082	23502	06136	89476
81686	51330	58828	74199	87214	13727	80539	95037	73536	16862
79788	02193	33250	05865	53018	62394	56997	41534	01953	13763
92112	61235	68760	61201	02189	09424	24156	10368	26257	89107
87542	28171	45150	75523	66790	63963	13903	68498	02891	25219
37535	48342	48943	07719	20407	33748	93650	39356	01011	22099
95957	96668	69380	49091	90182	13205	71802	35482	27973	46814
34642	85350	53361	63940	79546	89956	96836	91313	80712	73572
50413	31008	09231	46516	61672	79954	01291	72278	55658	84893
53312	73768	59931	55182	43761	59424	79775	17772	41552	45236
16302	64092	76045	28958	21182	30050	96256	85737	86962	27067
96357	98654	01909	58799	87374	53184	87233	55275	59572	56476
38529	89095	89538	15600	33687	86353	61917	63876	52367	79032
45939	05014	06099	76041	57638	55342	41269	96173	94872	35605
02300	23739	68485	98567	77035	91533	62500	31548	09511	80252
59750	14131	24973	05962	83215	25950	43867	75213	21500	17758
21285	53607	82657	22053	88931	84439	94747	77982	61932	21928
93703	60164	19090	63030	88931	84439	94747	77982	61932	21928
15576	76654	19775	77518	43259	82790	08193	63007	68824	75315
12752	33321	69796	03625	37328	75200	77262	99004	96705	15540
89038	53455	93322	25069	88186	45026	31020	52540	10838	72490
62411	56968	08379	40159	27419	12024	99694	68668	73039	87682
45853	68103	38927	77105	65241	70387	01634	59665	30512	66161
84558	24272	84355	00116	68344	92805	52618	51584	75901	53021
45272	58388	69131	61075	80192	45959	76992	19210	27126	45525
68015	99001	11832	39832	80462	70468	89929	55695	77524	20675
13263	92240	89559	66545	06433	38634	36645	22350	81169	97417
66309	31446	97705	46996	69059	33771	95004	89037	38054	80853
56348	05291	38713	82303	26293	61319	45285	72784	50043	44438

TABLE 1 (Continued)

93108	77033	68325	10160	38667	62441	87023	94372	06164	30700
28271	08589	83279	48838	60935	70541	53814	95588	05832	80235
21841	35545	11148	25255	50283	94037	57463	92925	12042	91414
09210	20779	02994	02258	86978	85092	54052	18354	20914	28460
90552	71129	03621	20517	16908	06668	29916	51537	93658	29525
01130	06995	20258	10351	99248	51660	38861	49668	74742	47181
22604	56719	21784	68788	38358	59827	19270	99287	81193	43366
06690	01800	34272	65479	94891	14537	91358	21587	95765	72605
59809	69982	71809	64984	48709	43991	24987	69246	86400	29559
56475	02726	58511	95405	70293	84971	06676	44075	32338	31980
02730	34870	83209	03138	07715	31557	55242	61308	26507	06186
74482	33990	13509	92588	10462	76546	46097	01825	20153	36271
19793	22487	94238	81054	95488	23617	15539	94335	73822	93481
19020	27856	60526	24144	98021	60564	46373	86928	52135	74919
69565	60635	65709	77887	42766	86698	14004	94577	27936	47220
69274	23208	61035	84263	15034	28717	76146	22021	23779	98562
83658	14204	09445	40430	54072	82164	68977	95583	11765	81072
14980	74158	78216	38985	60838	82806	49777	85321	90463	11813
63172	28010	29405	91554	75195	51183	65805	87525	35952	83204
71167	37984	52737	06869	38122	95322	41356	19391	96787	64410
78530	56410	19195	34434	83712	20758	83454	22756	83959	96347
98324	03774	07573	67864	06497	20758	83454	22756	83959	96347
55793	30055	08373	32652	02654	75980	02095	87545	88815	80086
05674	34471	61967	91266	38814	44728	32455	17057	08339	93997
15643	22245	07592	22078	73628	60902	41561	54608	41023	98345
66750	19609	70358	03622	64898	82220	69304	46235	97332	64539
42320	74314	50222	82339	51564	42885	50482	98501	00245	88990
73752	73818	15470	04914	24936	65514	56633	72030	30856	85183
97546	02188	46373	21486	28221	08155	23486	66134	88799	49496
32569	52162	38444	42004	78011	16909	94194	79732	47114	23919
36048	93973	82596	28739	86985	58144	65007	08786	14826	04896
40455	36702	38965	56042	80023	28169	04174	65533	52718	55255
33597	47071	55618	51796	71027	46690	08002	45066	02870	60012
22828	96380	35883	15910	17211	42358	14056	55438	98148	35384
00631	95925	19324	31497	88118	06283	84596	72091	53987	01477
75722	36478	07634	63114	27164	15467	03983	09141	60562	65725
80577	01771	61510	17099	28731	41426	18853	41523	14914	76661
10524	20900	65463	83680	05005	11611	64426	59065	06758	02892
93185	69446	75253	51915	97839	75427	90685	60352	96288	34248
81867	97119	93446	20862	46591	97677	42704	13718	44975	67145
64649	07689	16711	12169	15238	74106	60655	56289	74166	78561
55768	09210	52439	33355	57884	36791	00853	49969	74814	09270
38080	49460	48137	61589	42742	92035	21766	19435	92579	27683
22360	16332	05343	34613	24013	98831	17157	44089	07366	66196
40521	09057	00239	51284	71556	22605	41293	54854	39736	05113
19292	40078	06838	05509	68581	39400	85615	52314	83202	40313
64138	27983	84048	42635	58658	62243	82572	45211	37060	15017