NATIONAL OPEN UNIVERSITY OF NIGERIA

SCHOOL OF MANAGEMENT SCIENCES

COURSE CODE: PSM 823

COURSE TITLE:
STATISTICS AND OPERATIONS RESEARCH

COURSE GUIDE

PSM 823
STATISTICS AND OPERATIONS RESEARCH

Course Team Dr. O.I. Olateju (Course Developer/Writer) - LASU
Dr. Onwe (Course Editor) - NOUN
Dr. C. I. Okeke (Programme Leader) - NOUN
Mrs. P.N. Ibeme (Course-Coodinator) - NOUN

NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
14/16 Ahmadu Bello Way
Victoria Island, Lagos

Abuja Office
5, Dar es Salaam Street
Off Aminu Kano Crescent
Wuse II, Abuja

e-mail: centralinfo@noun.edu.ng
URL: www.noun.edu.ng

Published by
National Open University of Nigeria

Printed 2013

ISBN: 978-058-988-8

Printed by:

CONTENTS **PAGE**

INTRODUCTION

This course is a core course, which carries two credit units. The course material is prepared for all students who are taking any course in social and management sciences. The course will be a useful material to you in your academic pursuit, as well as in your work place as a manager and as an administrator.

WHAT YOU WILL LEARN IN THIS COURSE

The course is made up of 15 units, covering areas such as the nature of statistics, sampling techniques, organisation and presentation of data measures of distribution, probability and probability distribution. The course material also covers areas such as test of hypothesis, correlation and regression analysis.

The last part of the course material covers operation research. The topics covered include introduction to operational research, linear programming, transportation problem, games theory, network analysis and simulation.

COURSE AIMS

The main aim of the course is to expose you to the meaning and application of statistics. The course also aims at pointing out techniques of operation research that are needed for maximising organisational benefit and minimising cost of production.
The aim of the course will be achieved by:

- explaining the nature of statistics
- describing the different sampling techniques
- explaining and illustrating the different graphs and charts used in organising and presenting data
- identifying the different methods used for measuring central tendencies and measures of dispersion
- discussing the nature of probability and probability distribution
- explaining the procedure for hypothesis testing
- Describing the nature of correlation, regression analysis and operations research.

COURSE OBJECTIVES

At the end of this course, you should be able to:

- describe the nature of statistics
- state the different sampling techniques
- describe the graphs and charts used for organisation and presentation of data
- explain the different methods of averages used for measuring central tendency
- discuss the different methods used for studying dispersion
- highlight how linear programming, transportation problem, games theory, network analysis and simulation are used for optimising and reducing cost of production in organisations.

WORKING THROUGH THE COURSE

In this course, you will be exposed to the nature of statistics, the collection of data and how data are organised and presented. The course highlights the different sampling techniques. The course also examines the different methods of computing average and the method of studying their variation. The nature of probability and probability distribution are also highlighted in this course.
Other issues examined in this course include parametric statistics involving testing of hypothesis; methods of studying correlation and regression analysis and some of the techniques used for optimising and reducing cost of production in organisations.

COURSE MATERIALS

The major components for the course are as listed below.

(a) Course guide
(b) Study units
(c) Textbooks
(d) Assignments

STUDY UNITS

There are 15 units in this course which should be studied carefully. They are as follows:

Module 1

Unit 1 Nature of Statistics

Unit 2 Sampling Techniques
Unit 3 Organisation and Presentation of Data

Module 2

Unit 1 Measure of Central Tendency
Unit 2 Measures of Distribution
Unit 3 Probability
Unit 4 Probability Distribution

Module 3

Unit 1 Test of Hypothesis
Unit 2 Correlation and Regression Analysis
Unit 3 Analysis of Variance (ANOVA)
Unit 4 Analysis of Covariance (ANCOVA)

Module 4

Unit 1 Introduction to Operation Research
Unit 2 Linear Programming
Unit 3 Transportation Problem
Unit 4 Games Theory
Unit 5 Simulation
Unit 6 Network Analysis

THE ASSIGNMENT FILE

There are many assignments in this course and you are expected to do all of them by following the schedule prescribed- in terms of when to attempt them and submit same for grading by your tutor.

TUTOR-MARKED ASSIGNMENT

In doing the tutor- marked assignment, you are to apply and make use of the knowledge you have learnt in the contents of the study units. These assignments are expected to be turned in to your tutor for grading. They constitute 30% of the total score for the course.

FINAL EXAMINATION AND GRADING

At the end of the course, you will write the final examination. It will attract the remaining 70%. This makes the total scores to be 100%.

SUMMARY

This course exposes you to the nature of data presentation and analysis. On successful completion of this course, you will have been armed with the knowledge necessary for data analysis and techniques used for optimising and reducing cost in organisations.

MAIN COURSE

CONTENT **PAGE**

MODULE 1

UNIT 1 NATURE OF STATISTICS

CONTENTS

1.0 INTRODUCTION

Statistics is an intensive field of study which stems from the study of probability, resulting from information gathering. Therefore, in this unit, you will learn how you can use statistics to investigate and analyse data. You will, surely find this to be a rewarding experience.

2.0 OBJECTIVES

At the end of the unit, you should be able to:

- define statistics
- state basic statistical terms
- describe data collection processes
- explain how to write a good questionnaire.

3.0 MAIN CONTENT

3.1 Definitions

Here, let us look at the following concepts.

Statistics- this is concerned with scientific methods and theories which deals with data collection, organisation, summarisation, presentation, analysis, interpretation and utilisation of the results to draw useful conclusions, in order to make reasonable decisions.

Statistic- this is a numerical quantity whose values can be gotten from a sample data e.g. mean, median and mode of a data.

Data- these are facts, bits of information or series of observations that are obtained from an investigation. These facts can be measured or quantified.

Information- this is knowledge given. It is, necessarily, obtained from investigation or observation, and cannot always be measured or quantified.

Data set- these are data collected from a particular investigation or study.

Statistical data- these data are bits of information or measurements that are collected in the process of an investigation or observation.

3.2 Classification of Data

Data can be classified into two.

1. Numerical data
2. Non-numerical data

1. **Numerical data**- numerical data consist of values that can be quantified. They can be divided into discrete continuous data.

Discrete data- these are numerical data that consist of values that cannot assume values between two given values. They can only assume a particular value. Discrete data are whole numbers which can be positive, negative or zero integers, but does not include factions or decimal numbers. For example, 'the number of books in the book shelves can either be 0, 1, 2…and not 3½ or 4.6 $5\sqrt{3}$'.

Continuous data- these are numerical data which can assume values between two given values. They can be positive integers, negative integers or zero, fractions or decimals. Example of continuous data includes weight of items, height of pupils in a school age of individuals. In general, measurement gives rise to continuous data, while enumeration or counting gives rise to discrete data.

2. **Non-numerical data**- these are data which the value cannot be quantified. Examples are skin colour, nationality, gender, marital status, income group etc.; non- numerical data can also be divided into ordinal and categorical data.

Ordinal data-these are non-numerical data which the values can be arranged in an ordinal scale. They cannot be measured in natural numerical scale. Examples are income group, age group, classmate, boxers' weight, and academic grades. Academic grade is a label (an "A" grade student) that also belongs to a ranking system ("A" is better than "B")

Categorical data-these are non-numerical data which the values can only be put in categories. They cannot be arranged in an ordinal scale, but their measurement can be recorded on a natural occurring numerical scale. Examples are marital status, nationality, gender, religion, etc.

Variable-this is a quantity that varies, the opposite of a constant e.g. the numbers of customers that enters a shopping mall per hour varies, but the number of minutes per hour is constant.

Value-a value is a specific amount that a variable could be; for example- 'the number of callers to a call centre, in the last one hour, is 40'.

Observation or observed values- this is a value of a variable that has actually occurred or has been counted or measured.

3.3 Types of Statistics

Statistics, by nature, can be divided into two groups:

1. descriptive or deductive statistics
2. inductive or inferential statistics.

Descriptive statistics- this can also be referred to as deductive statistics; and it is the phase of statistics which seeks only to describe and analyse a given group without drawing any conclusions or inferences about a large group (population). Descriptive statistics covers areas such as

mean, median, mode, standard deviation etc., and can be presented in form of graph, data, tables, charts and other statistical tools. Example- 'the classification of the graduates of *PGD* students of Chartered Institute of Shipping (CIS) according to their sex, courses and *G.P.A.*, without drawing any conclusion or inference about the graduates is descriptive statistics.

Inductive statistics- this can also be called inferential statistics and it is that phase of statistics which enables one to draw conclusions or inferences about a large group, (population) from the analysis of sample drawn from the population. Most often, it is cumbersome and almost impossible to observe or collect data from the entire population; as a result, a part of the population- known as a representative sample, is considered, which represents the characteristics of the population. If the sample drawn is a true representative of the population, upon analysis of the observed sample, one can make inference about the population. The condition under which such inference is valid is inductive or inferential statistics.

3.4 Use of Statistical Data

Statistics plays a very vital role in day-to-day activities of man. Every field of life- ranging from business, account, banking, insurance, agriculture, energy, industry, science, engineering to medicine, make use of statistics. Some of the ways statistics has been found to be useful are as listed below.

1. It is used for effective budgeting, planning and forecasting future growth
2. It is used to summarise large data into concise information
3. It is used for the purpose of estimation and prediction of government revenue and expenditure
4. It is used for formulating policies for developmental purpose
5. The concepts, techniques and results of statistics is useful in modern day business- for good management decisions and ,effectively, planning future growth
6. It is used for both short and long-term forecast of future events
7. It is used in industries to control production process
8. It is used for making conclusions from data collected in experimental, social and behavioural research
9. Organisations use statistics to evaluate and monitor performance
10. It plays an important role in the usage of computers.

3.5 Sources of Data Collection

Sources of data, among numerous others, include the following.

1. **Agricultural statistics and data**- this contains agricultural statistics and data.
2. **Crime statistics and data**- this provides varieties of criminal justice statistics and data archives.
3. **Demographic data-** this deals with population, fertility and mortality data.
4. **Economic growth data-** this data is useful in conducting studies of economic growth.
5. **Educational statistics and data-** this deals with high school and college environments, educational cost, public sector investment and related information.
6. **Questionnaires-** these are popular means of collecting data, where questions are designed to get responses from respondents and the result of the responses are coded, quantitatively, for the purpose of analysis and interpretation.
7. **Health data**- this provides information on health care etc.

3.5.1 Methods of Collection of Data

Data collection can be seen as a means of gathering or obtaining information from selected sample of investigation. Without data collection, analysis cannot be done, and hence, meaningful planning and decision-making will be impossible. Therefore collection of data serves as required fact finding for any statistical analysis. Data collection enables us to evaluate numerical data and verify claims made about such data. It also helps in measuring the reliability of our inferences. There are six methods of data collection, as itemised below.

- Experimental method
- Direct observation
- Personal interview
- Telephone survey
- Registration
- Questionnaire

Experimental method- this is the method used in science laboratory to collect information. Here, data is obtained from laboratory experiments, and the results are recorded immediately.

Direct observation: this is a method of data collection, whereby the investigator watches and records what actually happens. The data collected here are more reliable and dependable. Data collected in this

manner does not give information about the past or future; it only gives information about the present.

Personal interview- this is a method whereby the interviewer and the respondent (interviewee) have a direct contact during the process of interview. Direct contact can be in form of face-to-face, one-on-one, 'many-to-one, or one-to-many.

Advantages

1. The interviewer can re-frame and explain unclear questions to the interviewee.
2. The data collected are original and direct from source.
3. It reduces ambiguity.

Disadvantages

1. It is expensive and time consuming.
2. There is the possibility of personal bias.

Telephone survey method- this is an oral interview via the telephone; this method is easy and convenient, but does not put into consideration those respondents that do not have access to telephone.

Advantages

1. It is convenient and fast
2. It has a wide coverage.

Disadvantages

1. It is expensive and exclusively for the rich.
2. The respondent may refuse to respond to the interviewer.

Registration method- this is the method of data collection whereby information is collected through registration. Example of data collection by registration method includes birth registration, marriage registration, and death registration.

Advantages

1. It is easy and saves time.
2. It reduces cost, especially, in carrying out census.
3. It gives estimate of population.

Questionnaire method- this is a method whereby a carefully, well organised questions- on all the information needed for an investigation, are written or printed and distributed to intended respondents for their responses. There are two types of questionnaires, namely:

i. Self-administered questionnaire
ii. mail/postal questionnaire

Self administered questionnaire-this is a method of data collection which involves distribution of questionnaire to the respondents to obtain the needed information, immediately.

Mail/postal questionnaire-this involves sending the questionnaire by mail/post to respondents. This technique often covers a wild geographical area, and it is mostly used when respondents are scattered. However, this method is not as effective as self-administered questionnaires, in that the respondent may refuse to send back the questionnaire to the interviewer; and the questionnaire could also be lost in transit.

Advantages

1. It is cheap.
2. It provides room for privacy as the respondent feels free to answer some questions.
3. It covers wide geographical area.
4. It is free of any possible bias from the interviewer.

Disadvantages

1. There could be loss of questionnaire in transit.
2. Refusal to send back the questionnaire.
3. The respondent may not have the required information. This is highly possible in an environment with high illiteracy level.

Designing a good questionnaire - a good questionnaire must be designed in such a way that it will reflect all the information needed for the study at hand. A badly designed questionnaire may ruin a well planned survey. Most importantly, questions and answers are the means of communication between the investigator and respondents; hence, the respondent, therefore, remains the only source of information available for decision-making. As a result, there are a number of rules we must follow, as a guide, when designing a questionnaire; these are as follows.

1. The questions must be short, direct and easy to understand.
2. The questions must not be difficult or ambiguous.

3. Avoid questions that lead to a particular answer; e.g. ask- "which brand of beverage do you consider the best?" and not- "do you consider Milo to be the best brand of beverage?", mentioning a name can lead to a particular answer.
4. Ensure that the respondent has the information.
5. The questionnaire must be properly edited and be in order, so as to give room for completeness, consistency and homogeneity.
6. The questions must not be personal or offensive.
7. Consider if the respondent will be willing to tell the truth, if telling the truth can affect his personality and integrity.

3.5.2 Types of Data Collection

There are, basically, two main types of data collection, these are listed below.

- Primary data
- Secondary data

Primary data

Data is collected from a representative sample for a particular study or investigation, for a particular purpose by means of sample survey; data collected are used for the purpose for which they are collected. Primary data is the most advisable to use in statistical analysis, because of its accuracy, reliability and dependability in decision- making.

Advantages

1. The purpose of collection is as follows.
2. Sorting out is possible; that is, the part of the data that are not relevant to the study can be removed or ignored.
3. Sampling procedure is known.

Disadvantages

1. It is expensive to carry out.
2. It involves a lot of steps, especially, in a large survey.
3. Sorting out must be done before analysis.
4. Sorting out may be cumbersome, especially, in large survey.

Secondary data

This relates to data which already exists, and which is used for a purpose other than that for which it was collected. Secondary data are obtained from publications, journals, records of government, dailies, and news

magazines; agencies are engaged in routine data collections. These agencies include the following.

1. The federal and state office of statistics
2. Central Bank of Nigeria
3. Research institutes and universities
4. Industries and commercial organisations.

Advantages

1. It is cheap and easier to collect, because it is readily available.
2. It already existed, and it is sorted and processed.
3. It facilitates and speeds up analysis.

Disadvantages

1. Errors involved in the original collection cannot be detected and eliminated.
2. Its sampling procedure is not known.
3. It is less used, compare to primary data.
4. The purpose of collection is not known, and so it may not be suitable for the investigation.

3.5.3 Errors in Data Collection

1) **Ineffective communication between the investigator and respondent**- this relates to communication gap between parties involved.
2) **Problem of difficult questions**- as it has been discussed above, some questions on questionnaires may be too difficult and unclear for respondent to understand
3) **Problem of exaggeration on the part of respondents**- some respondents may be biased because of personal affiliation to the administrator or attendant or interview and therefore, play prank with responses.
4) **Financial and logistic problem**- the investigator may have challenges in the area of transportation and funding; this may affect data collection, which in turn, can affect the investigation negatively.

5) **Return of incomplete questionnaire- as respondent may skip some key questions**. When an investigator hurriedly administers questions on respondent, the respondent may fill that the best way to satisfy the curiosity of the investigator is to quickly tick his response, and in the process some questions are left unanswered.

6) **Respondent may not be willing to tell the truth-** there is an adage that says that "truth is always bitter"; respondents sometimes may be unwillingly to disclose the truth, especially in an interview situation. He/she may fill that saying the right thing at the time may be tantamount to divulging information that can implicate him/her.

Now, basically, these errors can be classified as shown below, as well.

a. **Sampling error**-the sampling techniques used may not valid so as to provide information regarding the population investigated.

b. **Transcription error**- an investigator may wrongly put down wrong data which could result to leverage points.

c. **Fe response errorals** - it is possible to give false response on questions that are not well understood.

d. **Measurement error**- it is most often observed that an investigator may use qualitative data for quantitative data. This kind of measurement will derail the authenticity of what is to be measured.

e. **Biased error**- the investigator may be judgmental. He may have concluded in his mind the result of the data prior to carrying out the investigation. Therefore, the sample used may not be a representative of the parent population.

f. **Incompleteness or non-coverage area required**- the design of the research work may be faulty and thus information regarding what is to be investigated may be lost.

g. **Compilation error**- even when the instruments for data collection are well planned in advance wrong compilation may entangle the result of the investigation.

h. **Time change error**- if the information or data collected is obsolete, it may not be useful for investigation. That is, the information may not be seasoned when current information surpasses it.

i. **Rounding error**- if digits or figures are rounded wrongly they affect the data collected.

j. **Secondary data error**- the errors involve in the collection of this kind of data may be taskful to detect since it does not originate from the researcher.

Now, from the above, you can infer that the following are the basic problems associated with data collection.

1. Ineffective communication between investigator and respondents.
2. Problem of difficult questions.
3. Problem of exaggeration on the part of respondents.
4. Financial and logistic problems related to data collection.
5. Return of incomplete questionnaire as respondent may skip some key questions.
6. Respondents may not be willing to tell the truth.

3.6 Essential Steps Involved in Sample Survey

The following are the steps to follow in conducting a sample survey.

1. Definition of aims and objectives of the survey
2. Literature search
3. Development or design of questionnaire- which will reflect all the information needed to achieve stated objectives
4. Determination of sample size and sampling method or technique
5. Total cost estimation
6. Pilot survey- a test run of the designed questionnaire for possible modification
7. Man power training
8. Main survey- this involves sending out trained personnel for data collection, using the validated instruments
9. Post estimation survey- to check the validation of the survey results and correct all possible errors
10. Survey result analysis and interpretation, using appropriate statistical techniques
11. Survey report and conclusion.

3.7 Steps Involved in Solving Statistical Problem

Here, take note of the following.

1. Define the problem
2. Search for cause of the problem
3. Provision of alternative solutions
4. Evaluation of the solution
5. Choose a cause of action and measurement criteria.

3.8 Essential Steps in Carrying Out a Statistical Research

You are to, equally, note the following basic steps.

1. Formulation of research question.
2. Choose the right statistical test for your research design.
3. Important issues such as design of questionnaire, ethics, sampling, reliability and validity should be given due consideration.
4. Conducting sample statistics to describe and summarise your data.
5. Using statistics to explore relationships and differences in your data.
6. Writing of research report.

SELF-ASSESSMENT EXERCISE

Outline 10 ways you can use statistics.

4.0 CONCLUSION

In this unit, you have learnt the best way to apply statistics in various disciplines. It has been mentioned to you that data collection process takes two forms, namely- primary data and secondary data. This part is very crucial to your research process; therefore, you can now go on to apply what you have learnt in this unit. Let us examine other aspect of statistics in the next unit which is also pertinent to your study of statistics.

5.0 SUMMARY

So far in this unit, you have learnt about the nature of statistics. Some vital areas of statistics have been exposed to you-these include data, information, categorical and non-categorical data and ways to write a good questionnaire. You can now proceed to apply all that you learnt here.

6.0 TUTOR-MARKED ASSIGNMENT

i. a. What is statistics?
   b. Distinguish between the following and give two examples of each.
   i. Numeric and non-numeric data
   ii. Discrete and continuous data
   iii. Ordinal and continuous data
   iv. Ordinal and categorical data

v. Data and data set

7.0 REFERENCES/FURTHER READING

Adamu, I. M. (2006). *Understanding Basic Statistics*. Nile Ventures.

Babatunde, L. & Elegbede, W. (2005). *Business Statistics: Concepts and Application*. (1st ed.). Life and Ministry Publications.

Kazmier, L. (1972). *Schaum's Outline of Theory and Problems of Statistics*. (SI Metric Edition). Singapore: McGraw Hill.

Lind, D., Marchal, W. & Wathen, S. (2010). *Statistical Techniques in Business and Economics*. (4th ed.). McGraw-Hill.

UNIT 2 SAMPLING TECHNIQUES

CONTENTS

1.0 INTRODUCTION

You have just learnt about data collection techniques in Unit 1. However, for you to have reliable and valid data there is need to understand the sampling procedure that can ensure better data processing. This is within the purview of sampling techniques.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- define some basic elements of sampling techniques
- explain types of sampling.

3.0 MAIN CONTENT

3.1 Population

A population is a set or collection of objects, units having similar, observable characteristics. A population may be finite or infinite.

Finite population- the units are countable with a frame i.e. a list of all the units.

Infinite population- this could be countable, but with an impossible frame.

3.1.1 Sample

A sample is any subgroup or subset of the population under consideration. It can simply be referred to as a part or fraction of a whole. Since it is often difficult and cumbersome to access the entire population, a sample or subset is taken to obtain information about the population. It is important to be able to infer the required information about a population from the sample drawn from it.

Sampling is the method or the procedure of collecting information from a population. It is difficult to enumerate an entire population, due to some constraints such as in- accessibility of the population, time frame factor, cost, inadequate resources etc. Based on all these factors, there is a need to take sample so as to minimise these constraints. In order for the result of the analysis of the sample to be reliable, the sample taken must be a representative sample of the population. Sampling can therefore be referred to as a process of obtaining representative sample from a population.

On the other hand, sampling frame is a list of all the members or items in a population from which a sample is to be drawn. A book that contains the list of all the pupils in a class of a school and the relevant details (class register) is an example of sampling frame.

3.1.2 Total Enumeration

This can also be referred to as compete enumeration or census. It is the complete counting of all the individuals in a target population together with other vital or important information about every element of the population. For example, a census of people in Nigeria consists of all individuals, children and adults and other vital information about sex, age, qualification, marital status, number of children, number of dependants, academic qualification, number of the employed and the unemployed etc. Total enumeration is always difficult to carry out, it is prone to mistakes, it is expensive and time consuming, but free from sampling error.

Advantages of samples over complete enumeration

1. **Economic importance** - it is more economical to draw samples which will serve as a true representative of a population, than to carry out complete enumeration of the population. Printed questionnaire to use will be less and the manpower required will be minimal.
2. **Time factor** - it saves time to take sample and enumerate. Complete enumeration is time consuming; and for the statistical

data to be useful and effective it must be readily available, within the time frame it is needed. Otherwise, the information provided may become less useful, outdated and invalid.

3. Where the population size is large and indeterminable, it is more appropriate to use samples, rather than complete enumeration because the population may be too large to cover.

3.2 Sampling Techniques

There are, basically, two types of sampling techniques-as shown below.

1. Probability sampling techniques
2. Non-probability sampling techniques

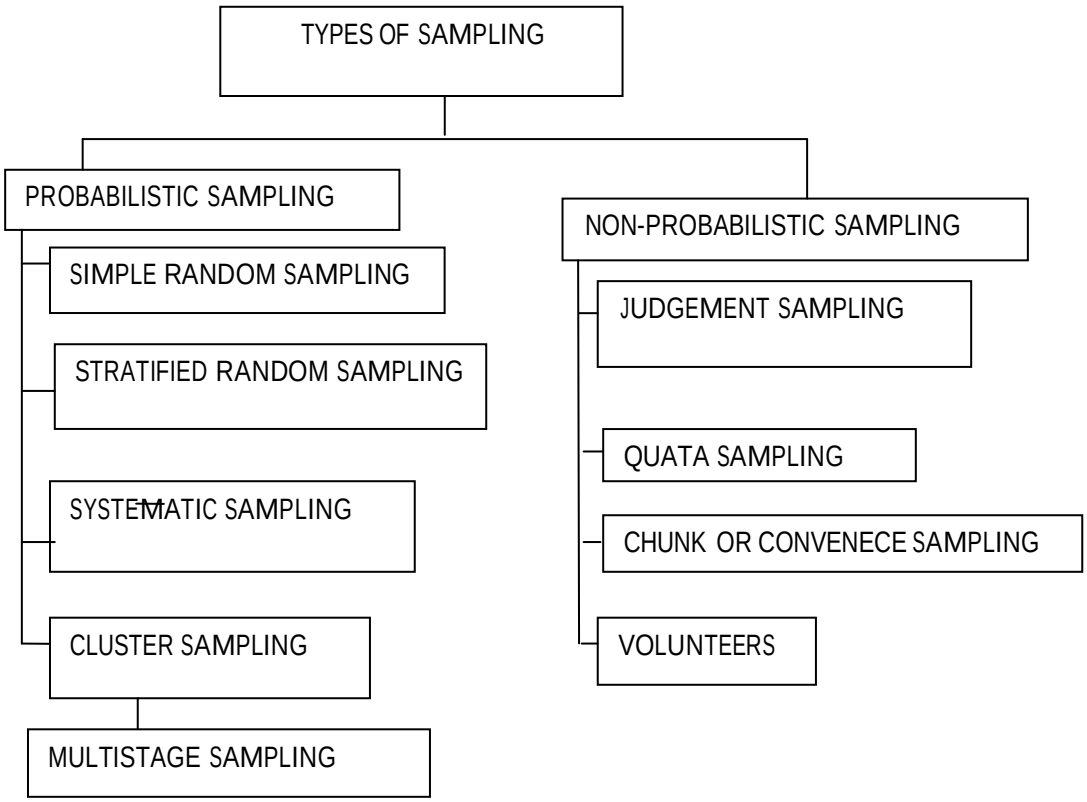


Fig. 2.1: Schema Showing the Typology of Sampling Techniques

3.2.1 Probability Sampling Technique

This selection is based on the use of existing probability law where each element in the population is known. These techniques include the list below.

a. Simple random sampling technique
b. Stratified random sampling technique
c. Systematic sampling technique

d. Cluster sampling technique
e. Multistage Sampling technique

Let us look at these one by one.

a. Simple random sampling technique

This is a sampling technique in which every possible sample (items) of size ***n*** taken from the population size *N* has equal and independent chances of being chosen. The selection is not biased. This sampling method or technique can also be referred to as unrestricted random sampling. To ensure randomness of selection, either of the following may be adopted.

i. **Lottery method**- this is a very popular way of taking a random sample. Number or name all the unit (item) of the population and place the folded slip of each unit of the population in a drum and mix thoroughly, then a selection is made of the number of slips required to constitute the desired size of sample(while the person doing the selection is a blindfolded).

ii. **Table random numbers**- here, random numbers are, generally, obtained by some mechanism, which when repeated a large number of times ensures ,approximately, equal frequencies for the numbers from 0 to 9 and also proper frequencies for various combinations of numbers (such as 001, 01, …..99; 000,001…999 etc) that could be expected in a random sequence of digits 0 to 9.

b. Stratified random sampling technique

This sampling technique is used when the population is divided into homogeneous groups or classes called strata; then simple random sampling technique is then used to draw the numbers of the sample from each of the stratum proportion of the size of that group, in the entire population. This type of technique is used when members of the population are from various economic and social groups. In this technique, the various groups forming the total population must be known and at what proportion.

Example

In a survey of 50 *PG* students of the Charted Institute of Shipping where the ratio of female to male is 10:40, draw a sample of 15 students using stratified sampling method.

Solution

Since the ratio is known, you can use stratified sampling technique-
N = population =50
Strata; $S_1 = 10$, $S_2 = 40$
Sample size = 15
Number of females= S_1/N x n = 10/50 x 15 = 3
Number of males= S_2/N x n = 40/50 x 15 = 12
Therefore, 3 female students will be chosen form 15 and, 12 male students will be chosen from 40- using simple random sampling technique.

c. Systematic sampling

Systematic sampling is a sampling technique where a complete list of the population from which sample is to be drawn is available and complete from the population. Select a random number between *1* and *k* in to the sample and every $\boldsymbol{k^{th}}$ element thereafter. Here, there is a need to take a decision, in advance, about the pattern to be used in selecting a sample from the population.
To select ***n*** samples out of a population of ***N*** where ***n*** is a factor of ***N***.

That is ***N= nk,*** you can take the following steps:

a. assign number *1-N* to every member of the population
b. identify the sampling ratio - K = N/n
c. select the first member by a simple random sampling method, between 1 and K and then select every $\boldsymbol{k^{th}}$ element after the random start
d. then *n* sample will appear.

Example

Use systematic sampling methods to select a sample of size 10 from a population consisting of 80 members. Given that the fist selected is 3(three).

Solution

Let x, x+k, x+2k, x+3k, ……. be the serial numbers
when x=3, then, k=N/n =80/10=8
Then serial numbers are 3, 3+8, 3+2(8), 3+3(8)……
=3,11,19,27,35,43,51,59,67,78.

Example

Below is the list of series made by twenty (20) students

Series No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Scores	41	70	30	0	90	65	56	33	45	70	55	57	38	49	6	13	22	91	69	80

Required

Take a sample of size 5 (five) to be systematic sampling, if the first observation in the sample is the fourth observation in the sample frame.

Solution

Unit to be included in the population are those with serial number to be determined by x, x+k, x+2k, x+3k…….

Where x = 4, k = N/n = 20/5 = 4

The serial numbers are: 4, 4+4, 4+2(4), 4+3(4)……
= 4,8,12,16,20.

Therefore, the sample of size 5 is 0,33,57,13, and 80.

d. Multistage sampling technique

In this sampling technique, the elements in the population are divided into small groups. At first, the first stage units are sampled by a suitable method such as simple random sampling. Then, a sample of second stage units is selected from each of the selected first stage units by some suitable method, which may be the same or different from the method used in the stage units. More stages may be added if required. For
Example - if Nestle Nigeria Plc wishes to know the perception of Nigerians about two of its products Nescafe and Milo; it may be difficult to conduct complete enumeration of the total population of the entire country. It is therefore expected that some cities such as Lagos, Ibadan, Abuja, Port-Harcourt and Kaduna will be taken as areas of consideration; then the sample for Nescafe and Milo can be drawn for those selected areas.

e. Cluster sampling technique

Here, the population is divided into groups that are not, necessarily, homogenous. The grouping is based on certain characteristics identified by the researcher; this can be geographical location, nearness to market etc. The sample is selected such that all the different clusters are represented in the sample.

3.1.2 Non-Probabilistic Sampling

These are sampling techniques that are not probabilistic in estimating and interpreting the representative sample from the population. Non-probabilistic sampling includes the following.

a. Judgement sampling or purposive sampling
b. Quota sampling technique
c. Chunk /convenience sampling
d. Volunteer sampling technique

a. Judgement sampling technique

This is a sampling method in which the choice sample is dependent, exclusively, on the discretion of the researcher. The researcher decides on what items to be included in the sample, based on what he thinks is most useful or needed.

b. Quota sampling technique

This is a form of judgment sampling. In a quota sampling, the population is broken down into groups based on some established characteristics related to the population such as income groups, age, sex, race, geographical location, etc.; each interviewer is then required to interview a certain number of persons which constitute the quotas. The selection of sample items depends on personal judgement.

c. Chunk or convenience sampling technique

This is a sampling method carried out with convenience. For example, if a researcher wishes to carry out investigation on the effect of motivation on performance in water producing company; if he/she decides to use a company close to his/her house, he/she has used convenience method of sampling- irrespective of the method of selection, as the company chosen may not be a true representative of the population.

d. Volunteer sampling technique

Here, sampling is done without any compulsion.

SELF-ASSESSMENT EXERCISE

How will you use systematic sampling to select a sample of size 15 from a population consisting 90 members?

4.0 CONCLUSION

This unit has enabled you to understand what you should do whenever you have chosen a research topic. You will discover that the method of sampling is paramount to the population size you have prescribed for your research work. It equally affects your experimental design.

5.0 SUMMARY

In this unit, you learnt the following:

- types of sampling techniques. You learnt about simple random sampling techniques- which means that every member of the population has to be included in the sample, and you were exposed to the method you can use to achieve this. You also learnt about quota sampling technique which is non-probabilistic. It is, especially, convenient for a wide spread geographical population.
- the importance of sampling techniques.

6.0 TUTOR-MARKED ASSIGNMENT

i. a. State the advantages of samples over total enumeration.
   b. List and explain the steps you would take in sampling survey.

7.0 REFERENCES/FURTHER READING

Adamu, I. M. (2006). *Understanding Basic Statistics*. Nile Ventures.

Babatunde, L. & Elegbede, W. (2005). *Business Statistics: Concepts and Application*. Life and Ministry Publications.

Kazmier, L. (1972). *Schaum's Outline of Theory and Problems of Statistics*. (SI Metric Edition). Singapore: McGraw Hill.

Lind, D., Marchal, W. & Wathen, S. (2010). *Statistical Techniques in Business and Economics*. (4th ed.). McGraw-Hill.

UNIT 3 ORGANISATION AND PRESENTATION OF DATA

CONTENTS

1.0 INTRODUCTION

In this unit, you will learn how you can process crude data or raw data into a refined one. The way you organise and represent your data, before analysis and interpretation, will impact, greatly, on the quality of your work. Therefore this unit will expose you to the pattern of data organisation and presentation.

2.0 OBJECTIVES

At the end of the unit, you should be able to:

- classify data and arrange
- illustrate the graphical work of classified data
- estimate median, mode, and other quartiles from classified data.

3.0 MAIN CONTENT

3.1 Arrangement of Data

Arranging or classifying data in a systematic manner, is the first vital step in transforming data into information. The most basic way to present a qualitative data is to tabulate it - to arrange it in the form of a summary table.

Classification of data- this is the grouping together of data with same identified common characteristics of features. Classification can be done through any one, or combination of any of the following methods.

i. Classification according to chronology (time)
ii. Classification according to geographical location
iii. Classification according to quantity (e.g. students' grades are classified according to their scores or marks).
iv. Classification according to quality (e.g. workers are classified according to qualification/certification and experience).

3.2 Tabulation of Data

Tabulation is the arrangement and presentation of classified data into rows and columns. It helps to compress large mass of data to a distinct pattern that can easily be understood. There are, basically, two major types of tables, namely:

i. simple table
ii. complex table

i. Simple table

A tabular structure with two columns and many rows is referred to as a simple table; for example, consider the Nigerian Railway Corporation Grant Aided Revenue Returns (Jan – Sept 1999) table below.

***Table* 3.1: Nigerian Railway Corporation Grant Aided Revenue Returns (Jan-Sept 1999).**

Source: Annual Report of Federal Ministry of Transport

Title	**Actual Revenue 1999 (₦)**
Passenger/coaching	131,330,777.00
Freight	352,379,131.00
Rent and leases	11,426,752.00
Miscellaneous	114,204,474.00

ii. Complex table

This is a table that is formed when either the *sub or the caption* is divided into two co-ordinate parts.

Table 3.2

Commodities	**Year**	
	1996	1997
Wheat	88,295	10360
Cement	36,337	63779
Ballast stones	-	385582
LPFO and Bitumen	10,024	15588
Container	2440	4482
Fertilizer	11322	-
Others	69243	436991

3.2.1 Characteristics of a Table

A good table must have the following properties.

i. *Title*- a table must have a title; a title gives a description of the content of the table. The title should be short, brief, concise and comprehensive enough to depict the content of the table
ii. *Caption*- these are the heading for columns, and should be clearly stated.
iii. **Subs**- these are the heading of rows; if they are in numerical classes, they should be arranged in such a way that overlapping is avoided, and there should not be gap between classes.
iv. *Unit of measurement*- this indicates the unit of measurement of a particular item e.g. *Kg* for net weight, tons for quantity, naira for currency etc.
v. *Footnote*- this provides brief information about any part of the table that is not self-explanatory.
vi. *Source*- the source of the information must be specified and stated below the table. This may sometimes be the footnote.

3.3 Frequency Distribution

A tabular arrangement of data by classes- together with the corresponding class frequencies is called a frequency distribution or frequency table.

Tally method

This method is used to tabulate the class frequencies from data. Tallies are obtained by making strokes for corresponding scores. When four strokes have been made in a particular class, the fifth one will be a cross, over the first four, making it a bunch of five strokes- i.e. a value of five (5) tallies is denoted as ~~*IIII.*~~

a. Frequency distribution and histogram for discrete data

Here, let us consider the following example.

When a die was thrown forty times, the following results were obtained.

Table 3.3

3	5	3	2	4	2	5	4	1	6
5	1	2	1	3	2	4	2	1	2
2	4	1	3	2	2	4	2	3	2
3	2	1	3	3	3	3	1	6	5

Required

i. Construct a frequency distribution of the data using a tally method
ii. Construct a histogram for the data

Solution

Table 3.4

Score	Tally	Frequency
1	~~IIII~~ II	7
2	~~IIII~~ ~~IIII~~ II	12
3	~~IIII~~ ~~IIII~~	10
4	~~IIII~~	5
5	IIII	4
6	II	2

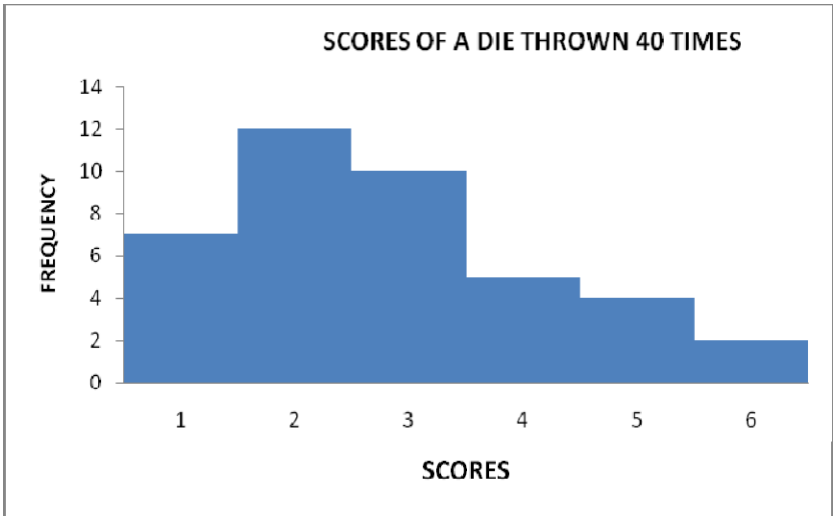


Figure 3.1

b. Frequency distribution and histogram for grouped data

Class - this is one of the categories into which quantitative data can be classified.

Class Frequency- this is the number of observations in the data set, falling in a particular class.

Example

The table below shows the frequency distribution of masses (recorded to the nearest Kg) of 100 male students in a university.

Table 3.5 Masses of 100 Male Students of Business Administration at Lagos State University.

MASSES (KG)	**NUMBERS OF STUDENTS**
50-52	8
53-55	15
56-58	40
59-61	28
62-64	9
	100

The first class or category, for example, consists of masses from 50 to 52 kg and is indicated by the symbol 50-52. Since eight students have masses belonging to this class, the corresponding class frequency is eight. Data organised and summarised as in the above frequency distribution are often called grouped data.

3.4 Class Interval and Class Limits

A category or class defined such as 50-52 in the above table is called a class interval. The end numbers, 50 and 52 are called class limits, the smaller number 50 is the lower class limit and the large number 52 is the upper class limit. The terms class and class interval are often used interchangeably, although the class interval is actually a symbol for the class. A class interval which has either no upper class limit or no lower class limit indicated is called an open class interval- e.g. referring to age groups of individual staff in a university, the class interval "60 year and above" is an open class interval.

a. Class Boundaries - if masses are recorded to the nearest *kg*, the class interval 50-52, theoretically, includes all measurement for 49.5000……..to 52.5000kg; these numbers indicate, briefly, by exact numbers 49.5 and 52.5, are called class boundaries or true class limits. The small number- 49.5 is the lower class boundary and the large number- 52.5 is the upper class boundary. In practice, the class boundaries are obtained by adding upper limit of one class interval to the lower limit of the next higher class interval and dividing by two. Sometimes, class boundaries are used to symbolise classes. For example, the various classes in the first column of the table 3.5 above can be indicated thus- 49.5-52.5, 52.5-55.5 etc; to avoid ambiguity in using such notation, class boundaries should not coincide with actual observations. Thus, if the value of an observation was 52.2, it would not be possible to decide whether it belonged to the class interval 49.5-52.5 or 52.5-55.5

b. The class or width of a class interval

The size or width of a class interval is the difference between the lower and upper class boundaries and is also referred to as the class width, class size, or class strength if all class intervals of a frequency distribution have equal widths, this common width is denoted by **c.** in such case c is equal to the difference between two successive lower limits or two successive upper class limits. For the data of table 2.1 for example, the class interval is

$C = 52.5 - 49.5 = 55.5 – 52.5$

c. The class mark

The class mark is midpoint of the class interval and is obtained by adding the lower and upper class limits and dividing by two. Thus, the class mark of the interval 50 – 52 is $(50 + 52)/2 = 51$. The class mark is also called the class midpoint.

For the purpose of further mathematical analysis, all observations belonging to a given class interval are assumed to coincide with the class mark. Thus all masses in the class interval 50 – 52kg are considered as 61kg.

3.5 General Rules for Constructing Frequency Distributions

Here, let us consider the following.

1. Determine the largest and smallest numbers in the raw data, and thus, find the range (difference between the largest and smallest numbers).

2. Divide the range into a convenient number of class intervals having the same size. If this is not feasible, use class intervals of different sizes or open class intervals. The number of class intervals is, usually, taken between five and 20, depending on the data. Class intervals are also chosen so that the class marks or midpoints coincide with actual observed data. This tends to reduce the so called grouping error involved in further mathematical analysis. However, class boundaries should not coincide with actual observed data.

3. Determine the number of observations falling into each class intervals, i.e. find the class frequencies using tally or score sheet method.

Example

The number of births recorded in each of the 50 maternity centres, in Surulere Local Government in October 2002, is as follows.

Table 3.6

50	99	81	86	69	85	93	63	92	65
77	74	76	71	90	74	81	94	67	75
95	81	68	105	99	68	75	75	76	73
79	74	80	69	74	62	74	80	79	68
79	75	75	71	83	75	80	85	81	62

Required

Construct a frequency distribution table using the interval-45-54, 55-64, etc.

Solution

Table 3.7

CLASS INTERVAL	TALLY	FREQUENCY	CLASS MARK	CLASS BOUNDARIES
45 – 54	I	1	50	44.5 – 54.5
55 – 64	II	2	60	54.5 – 64.5
65 – 74	~~IIII IIII IIII~~	15	70	64.5 – 74.5
75 – 84	~~IIII IIII IIII IIII~~ I	21	80	74.5 – 84.5
85 – 94	~~IIII~~ II	7	90	84.5 – 94.5
95 – 104	III	3	100	94.5 – 104.5
105 – 114	I	1	110	104.5 – 114.5
		50		

3.6 Histogram and Frequency Polygon

These are two graphical representations of frequency distributions

1. A histogram or frequency histogram consist of a set of rectangles having:
   a. bases on a horizontal axis (the X- axis) with centres at the class marks and lengths equal to the class interval sizes
   b. areas proportional to class frequencies. If the class intervals all have equal sizes, the height of the rectangles are proportional to the class frequencies and it is often customary to take the heights numerically equal to the class frequencies. If the class intervals do not have equal size, these heights must be adjusted.

2. A frequency polygon is a line graph of class frequency plotted against class mark. It can be obtained by connecting midpoints of the tops of the rectangles in the histogram

Example

The table below shows the number of child births recorded in each of the 50 maternity centres, in Surulere Local Government, in October 2002.

Table 3.8

Class Interval	Frequency
45 – 54	1
55 – 64	2
65 – 74	15
75 – 84	21
85 – 94	7
95 – 104	3
105 – 114	1
	50

Required

a. Construct a histogram for the distribution
b. A frequency polygon

Solution

Table 3.9

Class interval	Frequency	Class Mark	Class Boundaries
45 – 54	1	50	44.5 – 54.5
55 – 64	2	60	54.5 – 64.5
65 – 74	15	70	64.5 – 74.5
75 – 84	21	80	74.5 – 84.5
85 – 94	7	90	84.5 – 94.5
95 – 104	3	100	94.5 – 104.5
105 – 114	1	110	104.5 – 114.5
	50		

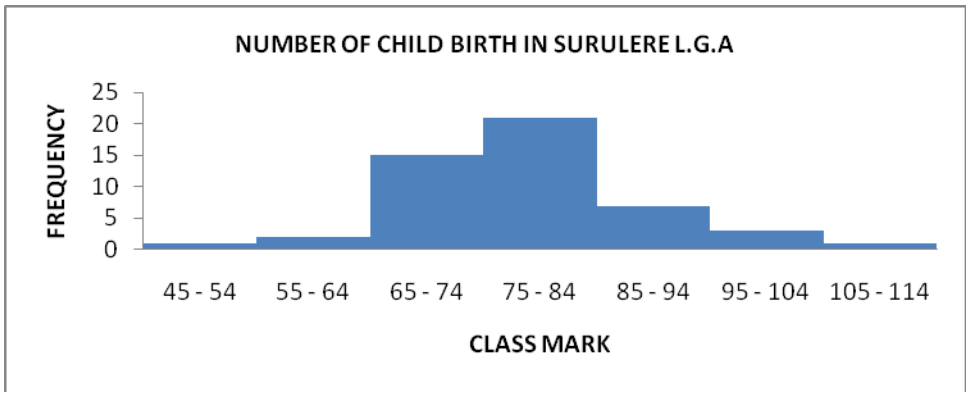


Figure 3.2

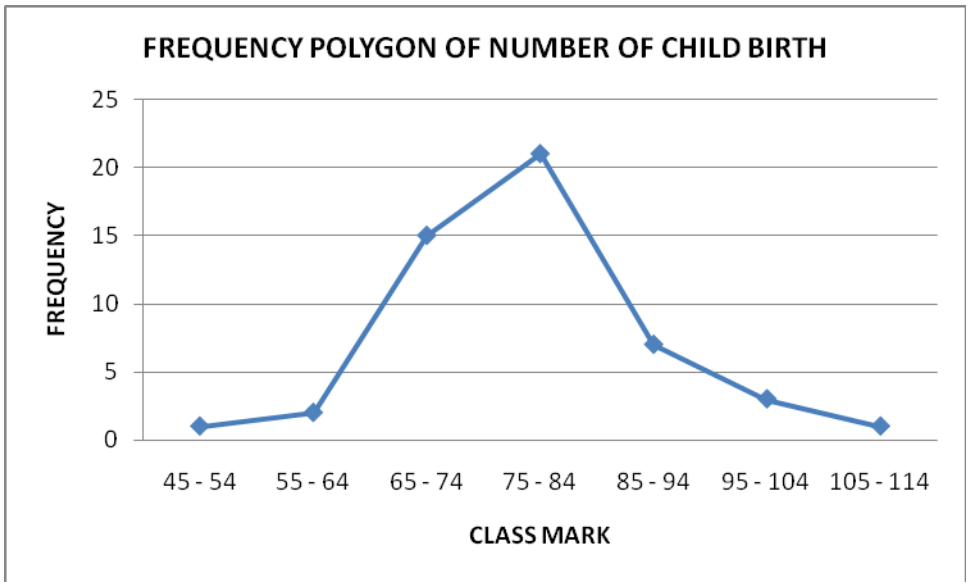


Figure 3.3

a. Relative frequency

The relative frequency is the proportion of observations within a category i.e. class frequency divided by the total number of observations in the data set.

$$\text{Relative Frequency} = \frac{\text{Frequency}}{\text{Sum of all Frequencies}}$$

Example

A physical therapist wants to determine the type of rehabilitation required by her patients. A random sample of the body part requiring rehabilitation is in the table below. Construct a relative frequency distribution.

Table 3.10

Body Part	**Frequency**	**Relative Frequency**
Back	12	12/30 = 0.4
Wrist	2	2/30 =0.0667
Elbow	1	1/30 = 0.0333
Hip	2	2/30 = 0.0667
Shoulder	4	4/30 = 0.1333
Knee	5	5/30 = 0.16667
Hand	2	2/30 = 0.0667
Groin	1	1/30 = 0.0333
Neck	1	1/30 = 0.0333

From the above table, you can see that the part of the body that requires rehabilitation is the *back.*

NOTE- check to ensure that sum of the relative frequency adds up to one. Sometimes it differs slightly due to rounding.

3.7 Presentation of Data

Data can be presented in charts, graphs and diagrams.

Charts and diagrams
Pictograms- this is the use of sample descriptive picture to represent data.
e.g., let ○ = 10 balls

Example

The table below shows the number of pupils from Christon Nursery and Primary School that passed common entrance examination to secondary school.

Year	**2007**	**2008**	**2009**	**2010**
No of pupils in a class	20	22	25	28

Required- represent the data, pictorially.

Solution

Let ○ = 10 balls

2007 ○○ = 20

2008 ○○◜ = 22

2009 ○○< = 25

2010 ○○◟ = 28

Pie charts

The pie chart is circular (spanning 360^0), and the size (angle) of the 'pie slice' assigned to each class is proportional to the overall total.

Example

Consider the following breakdown of how one family's income is spent in a month.

Table 3.11

	Amount (₦'000)
Mortgage and insurance	30
Electricity and gas	5
Food and drink	20
Clothes	10
Fares	8
Telephone	4
Car	15
Savings	5
Miscellaneous	3

Required-

Represent the above data in a pie chart.

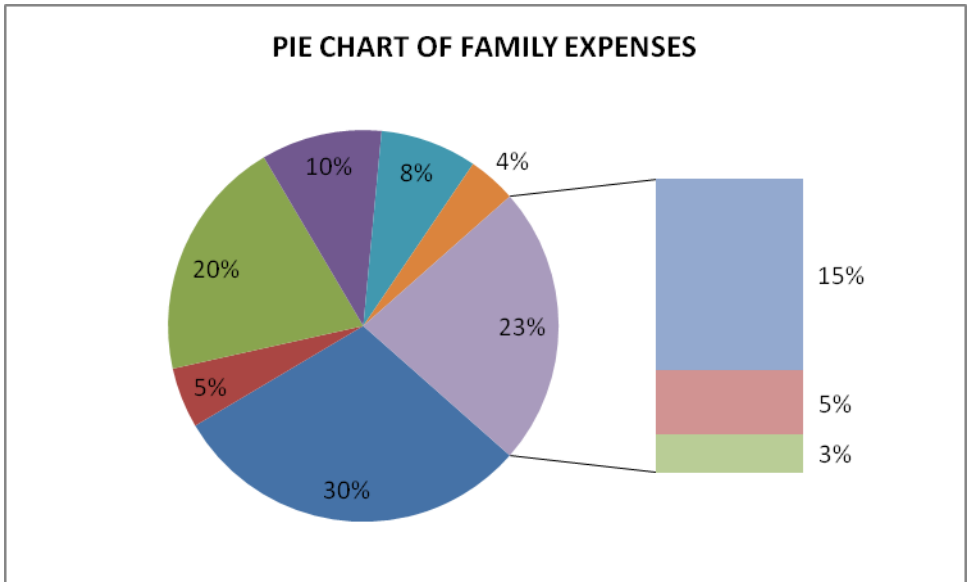


Fig. 3.4

Bar chart

Here, common examples are component bar chart, percentage bar chart, and simple bar chart. Simple bar chart is constructed by labelling each category of data on the horizontal axis, and the frequency or relative frequency of the category on the vertical axis. Rectangles of equal width are drawn for each category. The height of each rectangle is the frequency or relative frequency of the category.
Note- The bars do not touch each other.

Example

Skilodge com is test marketing its new website and is interested in how easy its web page design is to navigate. It randomly selected 84 regular internet users and asked them to perform a search task on the web page. Each person was asked to rate the relative ease of navigation as awesome, excellent, very good, good poor and very poor.

Table 3.12

Ease of navigation	**Number of workers**
Awesome	37
Excellent	17
Very good	11
Good	7
Poor	6
Very poor	6
Total	**84**

Required

Construct:

1. a bar chart for the data
2. a component bar chart
3. a compound chart

Note: Compound bar chart can also be referred to as component side-by-side chart or multiple bar charts.

Solution

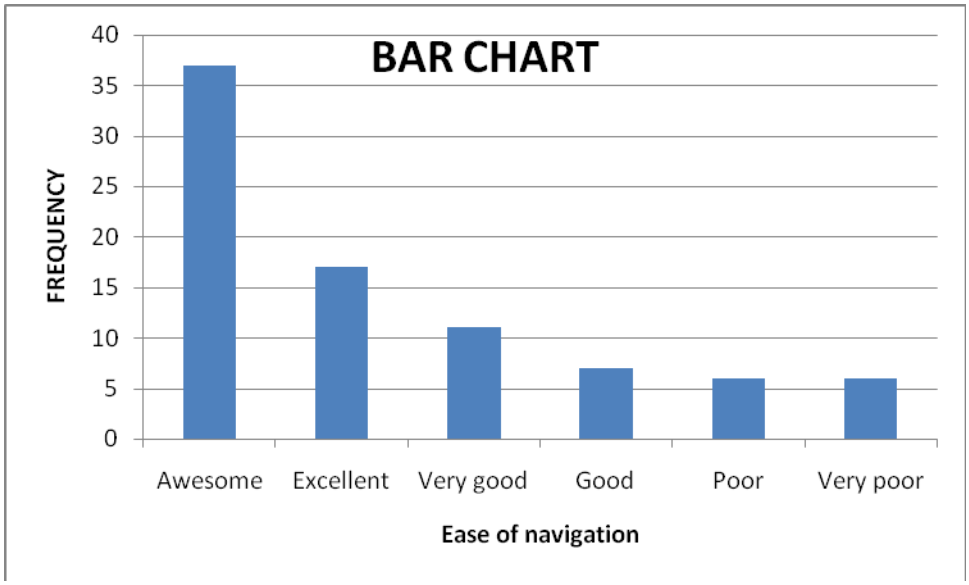


Fig. 3.5: Bar chart

SELF-ASSESSMENT EXERCISE

Briefly, describe various forms of bar diagrams.

3.8 Histogram of Distribution with Unequal Interval

When the frequency distribution of a given set of data are unequal, there is a need to compute an adjusted frequency such that the area of each bar will be proportional to its frequency and the total area of the histogram will be proportional to the total frequency. In order to compute the adjusted frequency, you are to identify the smallest class interval first, multiply it with the given frequency and then divide by the given interval

Note - if the interval of a group is double the smallest interval, divide the given frequency by 2, to obtain the adjusted frequency. Perhaps, if

the interval of a group is trice the smallest interval, then divide the given frequency by 3 to obtain the adjusted frequency etc.

Example

The monthly wages (in Naira) of 50 employees in a consulting firm are as shown below.

Table 3.13

Wages (N000)	**No of employees**
50 – 59	4
60 – 69	6
70 – 79	12
80 – 89	11
90 – 99	6
100 – 119	8
120 – 179	3
	70

Required

Construct a histogram for the frequency distribution

Solution

Table 3:14

Wages (N000)	Class boundaries	Frequency	Class interval	Adjusted frequency
50 – 59	49.5 – 59.5	4	10	10/10 X 4 = 4
60 – 69	59.5 – 69.5	6	10	10/10 X 6 = 12
70 – 79	69.5 – 79.5	12	10	10/10 X 12 = 12
80 – 89	79.5 – 89.5	11	10	10/10 X 11 = 11
90 – 99	89.5 – 99.5	8	10	10/10 X 8 = 8
100 – 119	99.5 – 119.5	6	20	10/20 X 6 = 3
120 – 179	119.5 – 179.5	3	60	10/60 X 3 = ½
		70		

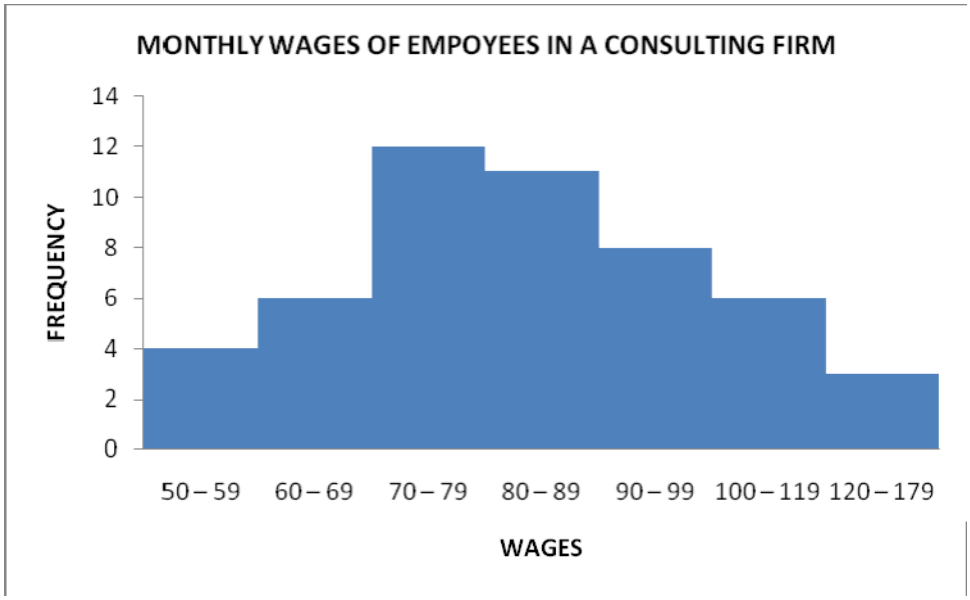


Fig. 3.6

3.9 Estimation of Median from Histogram

Median can be estimated by extrapolation- geometrically; median is the value of X (abscissa) corresponding to the vertical line which divides the histogram into two equal parts on equal areas. Since the area corresponds to frequency on a histogram, the value can be read, approximately- directly, from the histogram.

Example

The table below shows the examination score of 40 students in a class.

Table 3.14

Wages (₦'000)	Class Boundaries	Mid Point	Frequency
40 – 49	39.5 – 49.5	44.5	3
50 – 59	49.5 – 59.5	54.5	6
60 – 69	59.5 – 69.5	64.5	8
70 – 79	69.5 – 79.5	74.5	11
80 – 89	79.5 – 89.5	84.5	7
90 – 99	89.5 – 99.5	94.5	5

Required

i. construct the histogram
ii. determine the median from the histogram and explain how it is derived

Solution

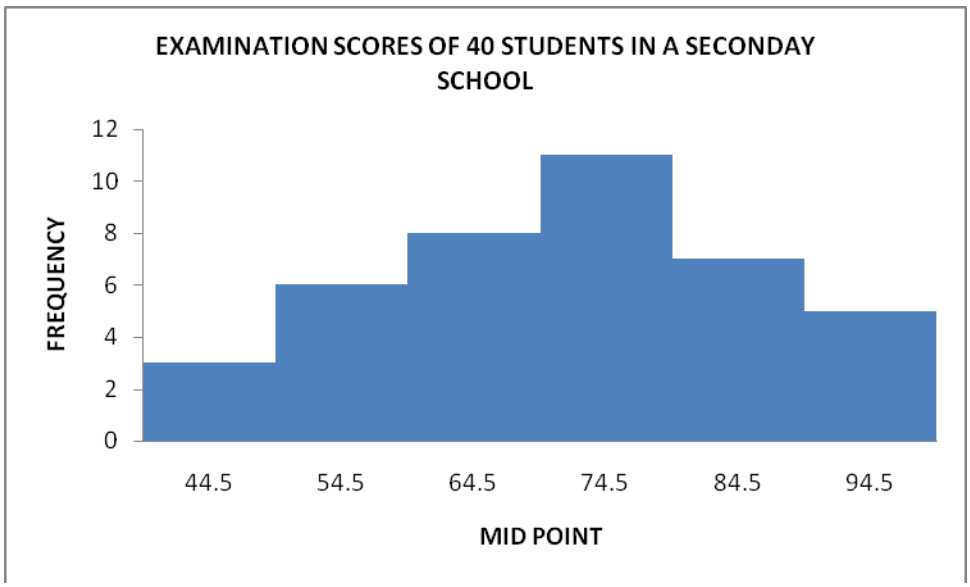


Fig. 3.7

Histogram

The median is the *X* axis (abscissa) corresponding to the line *PQ*, which divides the histogram into two equal areas. *PQ* is calculated such that the total areas to the right and left of it is half the total frequency- i.e. 20

The area *KQPR* and *OSNP* corresponding to frequencies of 3 and 8, respectively; then, we have KQ – 3/11 X KS =3/11 X 8 = 2.18. The corresponding class boundary of the median is 69.5. Therefore, the median = 69.5 + 2.18 = 71.68

3.10 Estimation of Mode from Histogram

Identify the modal class, the class just before the modal class and the class immediately after the model class. Draw a line from the top right of the modal class to meet to the top right corner of the class just before the model class; draw another line from the top left corner of the highest frequency to join the left corner of the class, immediately, after the modal class. The point where the two lines meet (intersect), draw a vertical dotted line to intersect the *X-axis* at point *NK*. Read the value of position *K*, and that is the mode.

Example

The following table indicates the mark scored by 100 students in quantitative techniques examination in a university.

Table 3.15

Marks	**Frequency**
25 – 34	3
35 – 44	10
45 – 54	21
55 – 64	33
65 – 74	20
75 – 84	8
85 – 94	5
	100

Required -

1. Construct a histogram for the frequency distribution
2. Estimate the mode from the histogram

Solution

Table 3.16

Wages (₦'000)	**Class Boundaries**	**Mid-Point**	**Frequency**
25 – 34	24.5 -34.5	29.5	3
35 – 44	34.5 -44.5	39.5	10
45 – 54	44.5 -54.5	49.5	21
55 – 64	54.5 -64.5	59.5	33
65 – 74	64.5 -74.5	69.5	20
75 – 84	74.5 -84.5	79.5	8
85 – 94	84.5 -94.5	89.5	5

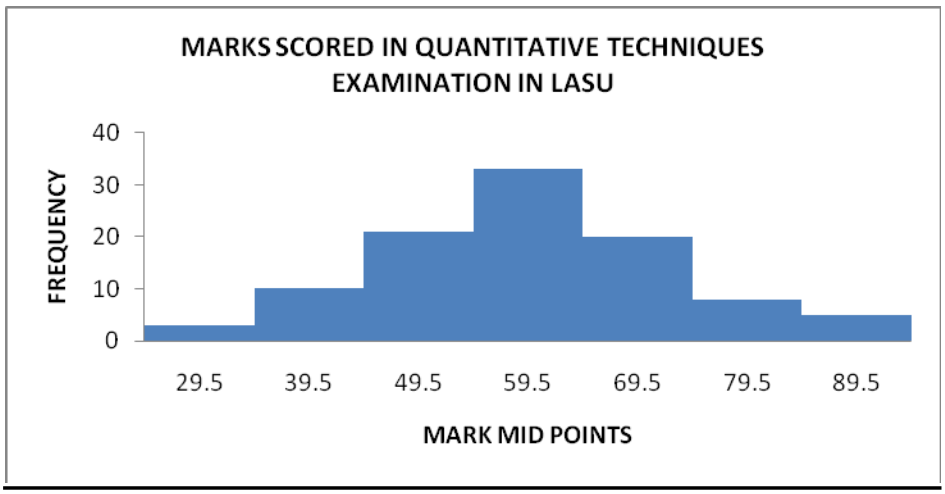


Fig. 3.8

4.0 CONCLUSION

You can now see that your ability to refine data is crucial. Having refined your data, you need to organise them to make them presentable. Various graphical works that will aid you have been provided in this unit. You should endeavour to apply them whenever you are pre-occupied with data organisation and presentation.

5.0 SUMMARY

In this unit, you have learnt how you can use histogram, pictogram, cumulative frequency curve, bar chart, frequency polygon etc., to give meaning to your data organisation and presentation. You are to apply this in your research work.

6.0 TUTOR-MARKED ASSIGNMENT

i. Draw a histogram, frequency polygon and frequency curve representing the following data.

Length of leaves (cm)	**Number of leaves**
6.5 – 7.5	5
7.5 – 8.5	12
8.5 – 9.5	25
9.5 – 10.5	48
10.5 – 11.5	32
11.5 – 12.5	6
12.5 – 13.5	1

ii. Given the data below, construct the O-give.

Marks	Number of students
0-9	9
10-19	42
20-29	61
30-49	140
40-49	250
50-59	102
60-69	71
70-79	23
80-89	2

7.0 REFERENCES/FURTHER READING

Otokiti, S., Olateju, O.I. & Adejumo, O. (2007). *Contemporary Statistical Methods*. (5th ed.). Lagos: Walex Printing Press.

MODULE 2 MEASURES OF CENTRAL TENDENCY AND VARIABILITY

UNIT 1 MEASURES OF CENTRAL TENDENCY

CONTENTS

1.0 INTRODUCTION

You have learnt much of data collection techniques, presentation and interpretation of data using descriptive statistics such as graphs (line graph, histogram, pie chart, bar chart, cumulative frequency curve etc.) and tables. However, you have not been exposed to interpretation of data by analysing them quantitatively. The quantitative measure of descriptive statistics, which is typically used by social and management science students, is at the core of measures of central tendency. Therefore, in this unit, you will learn some of the most commonly used measures of central tendency such as mean, mode, geometric mean etc.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- discuss the meaning and purpose of different measures of central tendency
- explain the different measures using ungrouped and grouped data
- illustrate the relationship between some of these measures of central tendency.

3.0 MAIN CONTENT

3.1 Arithmetic Mean

This is most popular measure of central tendency or measure of location; it is merely the average of the data. It is simply called the mean. Given a sample of data set, X_1, X_2 X_3……., X_n, the mean of this sample, denoted $\overline{X}$- (pronounced x bar), is given as-

$$\overline{X} = \frac{x_1 + x2\ldots\ldots + x_n}{n} = \frac{\Sigma xi,}{n}$$

Where- Σ (pronounced sigma is a shorthand notation for adding all data values given.)

Example 1

Given a sample of data such as 7, 4, 6, 5, 35, the mean which is denoted as-

$$\overline{X} = \frac{\Sigma xi}{n} \quad \overline{X} = \frac{7 + 4 + 6 + 5 + 3 + 5}{6} = \frac{30}{6} = 5$$

If each of the data value has frequency, the mean becomes-

$$\overline{X} = \frac{\Sigma xifi}{\Sigma fi}$$

3.2 Median

The median (which is often denoted as M_d) of a set of data is the middle value when the data are set in array from smallest to largest. For example, for the data above, the ordered array is- 3, 4, 5, 5, 6, and 7.
Thus, median (M_d) = $\frac{(n + 1)}{2}$*th* value. For odd data set, one value falls

in the middle; while for even data set two values fall in the middle- in such case, we usually add the values and divide by 2.

Example 2

From our first data set, the median becomes-

$M_d = \frac{(n+1)}{2}$ *th* value

i.e. $\frac{(6+1)}{2}$ *th* value $= \frac{(7)}{2}$ *th* value $= 3.5$ valueth. The array is

3, 4, 5, 5, 6, 7 and the median falls between 3rd and 4th values i.e.

$$\frac{5+5}{2} = 5$$

3.2.1 Mode

This is the value with the highest frequency or the number that occurs most in the data set. From the example above, the mode is 5, because it occurs most.

3.3 Geometric Mean

The geometric mean (denoted *g*) is used when arranging values to change. It is defined as:

$$g = \sqrt[n]{x_1 \text{ X } x_2 \text{ X } x_3 \text{ X } \ldots\ldots \text{ } x_n}$$

Where x1, x2, etc., represent rates of change between successive observations and - is the *nth* root (readily, calculated using calculator or computer). As a result of the cumbersomeness of its computation the geometric mean above can be simplified thus:

$$g = \sqrt[n]{x_1 \text{ X } x_2 \text{ X } x_3 \text{ X } \ldots\ldots \text{ } x_n}$$

$$= (x_1 \text{ X } x_2 \text{ X } x_3 \ldots \text{X } x_n)^{1/n}$$

Taking logarithm of both sides, we shall have-

$$\text{Log}(g) = {}^1/_n \text{Log } (\text{Log } x_1 + \text{Log } x_2 + x_3 + \ldots + x_n)$$

$$= {}^1/_n (\text{Log } x_1 + \text{Log } x_2 + x_3 + \ldots. + \log x_n)$$

$$= {}^1/_n \text{ } \Sigma \text{ Log } x$$

$$\text{Log}(g) = \frac{\Sigma \log x}{n}$$

$g = \text{Anti} - \log \quad \Sigma\left(\frac{\log_x}{n}\right)$

Where $X = x_1, x_2, x_3 \ldots\ldots x_n$.

Example 3

Find the geometric mean from the example above.

$g = \sqrt[6]{3 \times 4 \times 5 \times 6 \times 7 \times 5}$

$= \sqrt[6]{12{,}600}$

$= 4.83$

Or

Table 1.1

No	Log
3	0.4771
4	0.6021
5	0.6990
6	0.7782
7	0.8451
5	0.6990
	4.1005

$\therefore \quad \underline{4.005} = 0.6834$

6

$\text{Anti} - \log \quad 0.6834 = \underline{4.83}$

3.3.1 Trimmed Mean

This is the calculated mean without the highest and lowest values which are treated as outliers.

Example 4

Find the trimmed mean of the following data set. 1, 10, 12, 13, 40.

Solution

The lowest value is 1, while the highest value is 40. We will remove this, so that the remaining values are 10, 12, and 13.

$$\text{Trimmed} = \frac{10 + 12 + 13}{3} = 11.67.$$

3.3.2 Mid-Point

This is the average of the highest and lowest values of the data set.

Example 5

From example 1, the mid-point is

$$\frac{3+7}{2} = \frac{10}{2} = 5$$

3.3.3 Mid-hinge

This is the average of what is known as the first and third quartiles.

Example 6

From the data in example 1, calculate the Mid-hinge.

Solution

$$\text{Mid-hinge} = \frac{\frac{(n+1)th}{4} + \frac{3(n+1)th}{4}}{2}.$$

Example 7

Find the Mid-hinge in above example.

$$\text{First quartile} = Q3 = \frac{(n+1)th}{4} \text{ value}$$

$$= (^{7}/_{4})th \text{ value}$$

$$= \underline{0.175}$$

Arranging the numbers in array, we have 3, 4, 5, 6, 7,

$$Q1 = 0 + (3-0) \times 0.175$$

$$= 0.525$$

$$Q3 = \frac{3(n+1)th}{4}$$

$$= \frac{3(7)}{4} = 5.25 \; th \text{ value}$$

$$Q3 = 6 + 0.25 \quad (7-6)$$

$$= 6.25$$

$$\text{Mid-hinge} = \frac{0.525 + 6.25}{2} = \underline{3.388}$$

3.4 Harmonic Mean

This refers to the reciprocal of the arithmetic mean of the reciprocal of some given numbers. Given the data set - x_1, x_2 …. x_n , the harmonic mean *H*, is given as:

$$H = \frac{1}{\left(\sum_{i=1}^{N} \frac{1}{X_i}\right)}$$

$$H = \frac{N}{\sum \frac{1}{X}}$$

Example 7

Compute the harmonic mean in example.

$$H = \frac{N}{1/3 + ¼ + {}^1/_5 + {}^1/_6 + {}^1/_7} = \frac{6}{(0.33 + 0.25 + 0.2 + 0.2 + 0.17)}$$

$$= {}^6/_{1.29} = 4.65$$

3.5 Quadratic Mean or Root Mean Square

This refers to the square root of the arithmetic mean of their squares. This is denoted us R.M.S. Given the data x_1, x_2, x_3, ……..x_n

$$\text{R. M. S.} = \frac{\underline{x^2}_1 + \underline{x^2}_2, \underline{x^2}_3, \ldots\ldots\underline{x^2}_n}{N}$$

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

Example 8

Determine the R.M.S of example.

$$\text{R.M.S.} = \sqrt{\frac{7^2 + 4^2 + 6^2 + 5^2 + 3^2 + 5^2}{6}}$$

$$= \sqrt{\frac{49 + 16 + 36 + 25 + 9 + 25}{6}}$$

$$= 5.16$$

3.6 Weighted Mean (Average)

In order to give observation being averaged, their proper degree of importance, it is necessary to assign them (relative importance) weights, and then calculate a weighted mean. In general, the weighted average mean X_w of a set of data ($x_1 + x_2 + x_3 + \ldots + x_n$) with a relative importance that is weighted by a corresponding set of data ($w_1 + w_2 + w_3 + \ldots + w_n$) is given by

$$\overline{X}_W = \frac{\sum_{i=1}^{N} X_i W_i}{\sum W_i}$$

Example 9

In a business statistics course, the course assessment consists of homework 150, class work 450, project 100, and final exam 300. If 84% is homework, 97% is class work, 98% is project and 78% is final exam. Compute the weighted average.

Solution

$$\overline{X}_W = \frac{\sum_{i=1}^{N} X_i W_i}{\sum W_i}$$

$$= \frac{(0.84)(150) + (0.97)\,450) + 0.98\,(100) + (0.78)\,(300)}{150 + 450 + 100 + 300}$$

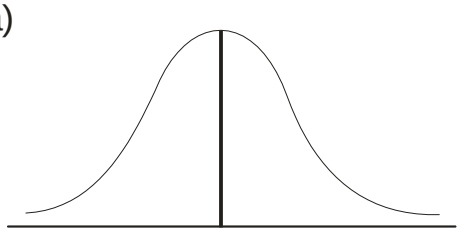
$$= 0.8845.$$

3.7 Relationship between Mean, Median and Mode

The figure below shows the relationship between the mean, median and mode for sets of data that are (a) symmetrical (b) positively skewed and (c) negatively skewed. It will be seen that the measures coincide in a symmetrical distribution but diverge in a non-symmetrical distribution. The mode is the same in all three cases but the mean is pulled in the direction of the skewness and the median falls in between the other two measures.

Symmetrical distribution or normal

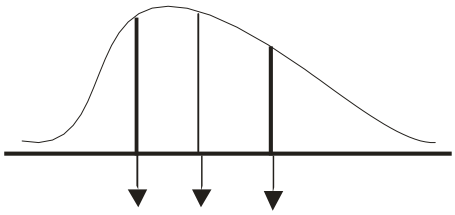
(a)



Mean
=
Median
=
Mode

(b)

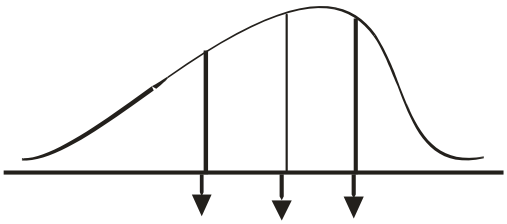
Skewed to the right.



Mode < Median < Mean
or
Mean > Median > Mode

(c)

Skewed to the left.



Mean < Median < Mode

or

Mode < Median < Mean

From the above, we can deduce that-

Median = $\frac{2 \text{ (mean)} + \text{mode}}{3}$

The mean and mode can be express in terms of median and mode and in terms of median and mean respectively.

3.8 Relationship between Arithmetic Mean, Geometric Mean and Harmonic Mean

The following are the relationships.

$X \geq g \geq H$ OR

$H \leq g \leq x$. They are the same if (and only if) the values of the data set are the same.

Example 10

Find the harmonic, geometric and the arithmetic mean of the data set 3, 3, 3, 3,

$$Hx = \frac{N}{\Sigma {}^{1}/_{x1}} = \frac{4}{½ + {}^{1}/_{3} + {}^{1}/_{3} + {}^{1}/_{3}} = \frac{4}{1.32} = 3.03$$

$$\overline{X} = {}^{12}/_{4} = 3$$

3.9 Grouped Data

So far, we have only dealt with raw and ungrouped data. However, a large set of data may be presented in grouped (or class) form as a frequency distribution. If this is the only data available for you to use, the calculation or the measure of location takes a different dimension. For grouped data therefore, the mean, median and mode are calculated as shown below.

Mean of grouped data

$$\overline{X} = \frac{\leqslant fx}{n}$$

Where x_i is the mid-point of the class intervals of *f*- the frequency of each class, and *n* data values

OR

$$\overline{X} = A + \frac{\leqslant Fd}{\leqslant f}$$

Where, *A* is the assume-mean, and d^1 is the coded factor used.

The median for grouped data must be found by interpretation. It is first necessary to identify the class in which the median value les and then to interpolate within this class as follows.

$$Md = L_{cmbd} + \left(\frac{{}^{N}/_{2} - CFBMC}{Fm}\right) \times W$$

Md = Median
L_{clmd} = Lower class boundary of the median class
N = Total numbers of observation
CFB_{ml} = Cumulative frequency of before the median class
Fm = Frequency of the median class
W = Class width

Mode of grouped data

For grouped data with equal class intervals (the mode *class I* -the class with the highest frequency), the mode is defined as follows-

$$M_0 = L_{cbmo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times W$$

L_{cbmo} = Lower class boundary of the modal class.

Δ_1 = (Pronounced Delta) = difference between modal frequency and the frequency immediately under it.

Example 11

Determine the mean, median and mode of the frequency distribution table below.

(Mean height of 50 Grade 10 Girls).

Table 1.2

HEIGHT (CM)	**FREQUENCY (F)**
150 – 155	4
155 – 160	7
160 – 165	18
165 – 170	11
170 – 175	6
175 – 180	4

Solution

(a) Mean = $\frac{\Sigma fx}{\Sigma f}$ or $A + \frac{\Sigma fd^1}{\Sigma f} \times c$

Height (cm)	**F**	**Midpt (X)**	**Fx**	**D = x – A**	**D1 = $^d/_c$**	**fd^1**	**Cmf**
150 - <155	4	152.5	610.0	-10	-2	-8	4
155 - <160	7	157.5	1,102.5	- 5	-1	-7	11
160 - <165	18	162.5	2925.0	0	0	0	29
165 - <170	11	167.5	1842.5	5	1	11	40
170 - <175	6	172.5	1,035.0	10	2	12	46
175 - <180	4	177.5	710.0	15	3	12	50
	Σf = 50		Σfx =8,225.0			Σfd^1 = 20	

$$\overline{X} = \frac{\Sigma fx}{\Sigma f} = \frac{8225.0}{80}$$

$$= 164.5\text{cm}$$

OR

$$\overline{X} = A + \frac{\Sigma fd^1}{\Sigma f} \text{ xc}$$

$$= 162.5 + {}^{20}/_{80} \text{ x } 5 = 164.5\text{cm}$$

(b) Median = Mr = $LCB_{Md} + \frac{({}^{N}/_{2} - CFB_{md})}{Fm}$ x w

$LCB_{Md} = 159.5$

$N = 50$, ${}^{N}/_{2} = 25$, $Fm = 18$

$CFBmd = 11$

$W = 5$

$$Md = 159.5 + \frac{25 - 11}{18} \text{ x } 5$$

$$= 159.5 + \frac{14}{18} \text{ x } 5$$

$$= 163.4\text{cm}$$

$$\text{Mode} = LCB_{M0} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \text{ x w.}$$

$\Delta 1 = 18 - 7 = 11$

$\Delta 1 = 18 - 11 = 7$

$LCB_{M0} = 159.5$

$$Mo = 159.5 + \frac{11}{11 + 7} \text{x } 5$$

$$= 159 - 5 + {}^{11}/_{18} \text{ x } 5$$

$$= 162.6\text{cm}$$

SELF-ASSESSMENT EXERCISE

A market – research firm conducted a household survey in Lagos metropolis. The following task shows the age distribution of the households.

Age (Years)	**No of Households**
10 – 14	4
15 – 19	8
20 -24	15
25 – 29	19
30 – 34	21
35 – 39	12
40 – 44	8
45 – 49	6
50 – 54	4
55 – 59	3

Compute: (a) the mean (b) the median (c) the mode.

4.0 CONCLUSION

In this unit, you learnt about measures of central tendency- used when analysing quantitative data. The arithmetic mean, median, mode and geometric mean are the four basic components of measures of central tendency or location; of these four, the best is the arithmetic mean. You shall later discover its usefulness in your study of variability and inferential statistics.

5.0 SUMMARY

In this unit, you have learnt another method which you could use to analyse quantitative data. You need to master the use of the formulae and apply them accordingly. Therefore, you are to note the following.

- Arithmetic mean- $\bar{X}_o = \frac{\sum Xi}{n}$

  For grouped data, it is computed using-

  $\bar{X} = \frac{\sum fiXi}{\sum f}$

- *Median*- a measure of central location or tendency of a data set; it is the value which splits the data set into two equal groups – one with values greater than or equal to the median and one with values less than or equal to the median (for ungrouped data). However, for grouped data the median is-Mo = LCBmc + $\left(\frac{N}{2} - \text{CMFBm}\right)$x c
  $\frac{\qquad\qquad}{\text{Fm}}$
- *Mode*- a measure of central tendency of a data set defined as the most occurring data value, in the case of ungrouped data. While for grouped data-

  $Mo = LCBm_{DC} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$

6.0 TUTOR-MARKED ASSIGNMENT

i. The following table lists the number of people killed in traffic accidents over a 10 year period. During this period, what is: (a) the average number of people killed per year? (b) the median for people killed per year? (c) the mode of people killed per year?
d. How many people died each day, on the average, during this period.

Year	1	2	3	4	5	6	7	8	9	10
Fatalities	959	1,033	960	797	663	652	560	619	623	959

ii. The following are the expenses on food and lodging incurred by a sample of 12 sales persons for the same week. For these data, compute: (a) the mean, (b) the median (c) the mode.

Sales man	1	2	3	4	5	6	7	8	9	10	11	12
Amount (N)	55	84	63	57	52	70	56	68	74	66	68	64

7.0 REFERENCES/FURTHER READING

Brian, P.M. & Philip, M.N. (2006). *Applied Statistics for Public Policy*. New Delhi: Prentice-Hall .

James, T. & Benon, P.G. (1988). *Statistics for Business and Economics*. (4th ed.). United States: Dellen Publishing Company Inc.

Okafor, R. (2004). *Statistical Methods Plus Non Parametric Techniques*. JAS Publishers.

UNIT 2 MEASURES OF DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

Another useful measure in quantitative data is the measures of spread. Already, you have learnt how to measure points; examples are mean, mode and median which best describes or represents the characteristics of the entire group. However, measures of dispersion or variability provide information that describes individual differences. In other words, what is the nature of the spread of the data around the mean or how much do the observations in a set of data vary from one another. This unit shall, therefore, provide you with information on modified ranges, standard deviation, variance, moment, skewness and kurtosis.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the meaning and purpose of different measures of variability
- calculate the different measures of variability using grouped and un-grouped data
- explain skewness and kurtosis.

3.0 MAIN CONTENT

3.1 Range

Range measures the spread of data by measuring the distance between the smallest and the largest measurements.

Example 1

Find the range of the following data set- 5, 96, 37

Range = largest - smallest
= 9 - 3
= 6

3.2 Modified Ranges

These are based on various measures which divide data set into equal parts such as quarters (quartiles), tenths (deciles) and hundredths (percentiles). Collectively, these are referred to as quartiles.

3.2.1 Quartile Measures

First (or lower) quartile Q_1, for un-grouped data is located at the $\frac{(n+1)}{}th$ observation-while for grouped data it 4 is located at the $(^{n}/_{4})th$ observation.

Third (or upper) quartile Q_3, for un-grouped data is located at the $3\frac{(n+1)}{4}th$ observation; while for grouped data, it is located at the $\frac{(3n)}{4}th$ observations, when N is the number of observations.

The second quartile (Q_2) is of course, the same as the median, which lies at the mid-point of the range of values. From the above, we can define the inter-quartile range and semi-inter-quartile range respectively.

Inter-quartile range- $Q_3 - Q_1$

Semi – inter-quartile range- $\frac{Q_3 - Q_1}{2}$

This is also referred to as quartile deviation.

Example 2

A rugby team scored the following number of points in each of their last ten matches respectively 18, 3, 21, 15, 9, 84, 27, 10, 42, 6; compute the following descriptive statistics of variability for the set of points:

(i) lower and upper quartile
(ii) inter-quartile range
(iii) semi-inter-quartile range

Solution

(i.a) Lower quartile Q1 = $\frac{(n + 1)}{4}$th

Given the following observations, arrange the numbers in ascending order of magnitude 3, 6, 9, 10,15, 18, 21, 27, 42, 84.

n = 10

$Q_1 = \frac{(10 + 1)}{4}$th $\frac{(11)}{4}$th observations.

$= 2.75^{th}$ observations.

6 + 0.75 (9 – 6)

6 + 0.75 (3) = 8.25

(i.b) Upper quartile $Q_3 = \frac{3(n + 1)}{4}$th $= \frac{3(14)}{4}$th $= \frac{(33)}{4}$th

Observations = 8 + 0.25 (42 – 27)

= 8 + 0.25 (15)

= 11.75.

(ii) Inter-quartile range = $Q_3 - Q_1$

= 11.75 - 8.25

= 3.5

(iii) Semi- inter-quartile range = $\frac{Q_3 - Q_1}{2}$

= $\frac{11.75 - 8.25}{2}$

= 1.75

Example 3

Given the table below, compute:

(i) lower quartile deviation (ii) upper quartile
(iii) inter-quartile range (iv) quartile deviation

Table 2.1

Age (Years)	No of Households
10 – 14	4
15 – 19	8
20 -24	15
25 – 29	19
30 – 34	21
35 – 39	12
40 – 44	8
45 – 49	6
50 – 54	4
55 – 59	3

Table 2.2

Age (Years)	No of Households	CMF
10 – 14	4	4
15 – 19	8	12
20 -24	15	27
25 – 29	19	46
30 – 34	21	67
35 – 39	12	79
40 – 44	8	87
45 – 49	6	93
50 – 54	4	97

Solution

(i) Q_1 = $(^n/_4)$th observation

= $\frac{(100)}{4}$th observation = 25 observations

This falls in cmf = 27

Q_1 = $LCBQ_1 + \frac{(^N/_4 - CMFBQ_1)}{FQ_1} \times W$.

$LCBQ_1$ = lower class boundary of the first quartile class.

First quartile class.

$CFMBQ_1$ = Cumulative frequency before the quartile class.

FQ_1 = Frequency or the quartile class.

Q_1 = $19.5 + \frac{(25 - 12)}{15} \times 5$

Q_1 = 19.5 + $^{13}/_{15}$ x 5

Q_1 = 23.8

(ii) Q_3 = $\frac{(3N)}{4}$th Observations

= $\frac{(300)}{4}$th observations.

= 75^{th} observations. It falls on CMF = 79

Q_3 = $34.5 + \frac{(75 - 67)}{12} \times 5$

= 37.8

(iii) Inter-quartile range $Q_3 - Q_1$

= 37.8 - 23.8

= 14.

(iv) Quartile deviation = $\frac{Q_3 - Q_1}{2}$ = $\frac{14}{2}$ = 7

3.2.2 Deciles

This is when data set is partitioned into ten equal parts.

Lower deciles (D_1) = $\frac{(n-1)}{10}$th observations for ungrouped data, while $(^{n}/_{10})^{th}$

Observations for grouped data. Upper deciles (D_9)

= $9\frac{(n+1)}{10}$th observations,

$\frac{(9n)}{10}$th observations for grouped data. Interdecite range therefore is defined as:

$D_9 - D_1$, while Semi-deciles range is defined as-

$$\frac{D_9 - D_1}{2}$$

Example 4

Use example 13 to compute: (i) lower deciles (ii) upper deciles (iii) Inter-decile range (iv) semi-inter-decile range

(i) D_1 = $\frac{(n + 1)}{10}$th observations

= $\frac{(10+1)}{10}$th observations = 1.1 observations

D_1 = 3 + 0.1 (6 – 3)

= 3.3

(ii) $D_9 = \frac{9(n+1)}{10} = \frac{9(11)}{10} = 9.9$ observations

$D_9 = 42 + 0.9\ (84 - 42)$

$= 79.8.$

(iii) Inter-decile range $= D_9 - D_1$

$= 79.8 - 3.3$

$= 76.5$

(iv) Semi – inter-decile range $= \frac{D_9 - D_1}{2} = 38.25$

Example 5

Use example 14, to compute: (i) lower deciles (ii) upper deciles (iii) inter-decile range (iv) semi – inter-decile range

(i) $D_1 = (^n/_{10})$th observations

$D_1 =$ 10 observations. This falls on cmf $= 12$

$D_1 = 14.5 + \frac{(10 - 4)}{8} \times 5$

$D_1 = 18.25$

$D_9 = \frac{(9N)}{10}$th observations $= 90$ observations

(ii) $D_9 = 44.5 + \frac{(90 - 87)}{6} \times 5$

$= 47$

(iii) Inter-decile range $= D_9 - D_1 = 28.75$

(iv) Semi-inter-decile range $= \frac{D_9 - D_1}{2} = 14.375$

3.2.3 Percentiles

This is the partitioning of a data set into hundredth equal part.

Lower percentile $(P_{10}) = \frac{(n + 1)}{100}$ 10th Observation

For ungrouped data while $\left(\frac{10nth}{100}\right)$ observations.

Upper percentile $(P_{90}) = \left(\frac{90(n+1)}{100}\right)$ the observations for ungrouped data.

While $90\ \frac{(n)}{100}$th observations for grouped data

Inter-percentile range $= P_{90} - P_{10}$

Semi-inter-percentile range $= \frac{P_{90} - P_{10}}{2}$

Example 6

Compute: (i) lower percentile (ii) Upper percentile (iii) inter-percentile range
and semi – percentile range, using example 13.

(i) Lower percentile $= P_{10} = \left(\frac{10(n+1)}{100}\text{th}\right)$ observations.

$= \frac{10(11)}{100}$

$= 1.1$ observation

$= 3 + 0.1\,(6 - 3)$

$= 3.3$

$P_{10} = D_1$

(ii) Upper quartile $= P_{90} = \frac{90(n+1)}{100}$th observations

$= \frac{90(11)}{100}$th observations

$= 9.9$

$= 42 + 0.9\,(84 - 42)$

$= 79.8$

$P_{90} = D_{10}$

3.3 Mean Deviation

The mean deviation of a set of data is the average of the deviations of every observation *Xi* from the mean *x,* and then taking the absolute value (II) of each of the data set. For instance, /-3/ reads- absolute value of -3 is + 3. Thus, mean deviation- sometimes called mean absolute deviation can be represented as follows.

MD or MAD $= \frac{\Sigma /xi - x/}{n}$ for ungrouped

and $f\,\frac{\Sigma/xi - x/}{\Sigma f}$ for grouped data

3.4 Coefficient of Quartile Deviation or Quartile Coefficient of Dispersion

A relative measure of dispersion based on the quartile deviation is called the coefficient of quartile deviation. It is defined as-

$$\frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

3.5 Coefficient of Mean Deviation

A relative measure of dispersion based on the mean deviation is called the coefficient of mean deviation or the coefficient of dispersion. It is defined as the ratio of the mean deviation to the average used in the calculation of the mean deviation. Thus-

Coefficient of M.D. (about mean) = $\dfrac{\Sigma/x - \bar{x}/}{n}$ / $\bar{x}$

Coefficient of M.D (about median) = $\dfrac{\Sigma / x - Md/}{n}$ / median

3.6 Coefficient of Dispersion

This is defined as the ratio of the MAD (Mean Absolute Deviation) to the median. It is measured as-

$$CD = \frac{MAD}{Md} = \frac{\Sigma/xi - Md?}{n} \Big/ Md$$

Example 7

The following are the number of miles between home and office of a sample of 10 clerical workers, employed by the same firm- 3,16,12, 12, 14, 5, 7,14,9,8.

For this data, find: (a) mean deviation (b) coefficient of quartile deviation (c) coefficient of mean deviation (about mean), (about mode) (d) coefficient of dispersion.

Solution

(a) $$MD = \frac{\Sigma/xi - x}{n}$$

$$\bar{X} = \frac{3 + 16 + 12 + 12 + 14 + 5 + 7 + 14 + 9 + 8}{10}$$

$$\bar{X} = \frac{100}{10} = 10$$

Table 2.3

X	$x - \bar{x}$	$1x - \bar{x}1$	x – Md(10.5)	/x – md/	/x – Mo/(12)	/x – Mo /
3	-7	7	-7.5	7.5	9	11
16	6	6	5.5	5.5	4	2
12	2	2	1.5	1.5	0	2
12	2	2	1.5	1.5	0	2
14	4	4	3.5	3.5	2	0
5	-5	5	-5.5	5.5	+7	9
7	-3	3	-3.5	3.5	+5	7
14	4	4	3.5	3.5	2	0
9	-1	1	-1.5	1.5	3	5
8	-2	2	-2.5	2.5	4	6
N =10		$\Sigma/x - \bar{x}/$ = 34		Σ/x – Md/ =36	Σ/x – Mo/ =36	Σ/x–Md/ =44

MD = 34/10 = 3.4

(b) C Q D (Coefficient of Quartile Deviation)

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Q_3 = (3(n +1)th observations
4

Arranging the data set in ascending order-3, 5, 7,8,9,12,12,14,14,16

Q_3 = 3(11)th observations = 8.25 observations
4

Q_3 = 14 + 0.25 (14 - 14)

= 14

Q_1 = (11)*th* observations = 2.75 observations
4

Q_1 = 5 + 0.75 (7 – 5)

= 5 + 1.5

= 6.5

CQD = $\frac{14 - 65}{14 + 6.5}$ = $\frac{7.5}{20.5}$ = 0.37

(Ci) Coefficient of deviation (about mean)

$$= Q_1 \quad \frac{\Sigma/x - Md/}{n}$$
x

= $\frac{3.4}{10}$ = 0.34

(ii) Coefficient of deviation (about median)

$$= \frac{\Sigma/X_i - Md/\,n}{md}$$

Md = (n +1)th value / 2

= (11/2)th values = 5.5^{th} values

Md = 9 + 0.5 (12 – 9)

= 10.5

Coefficient of deviation (about median) – from the table

$$= \frac{36}{10} = \frac{3.6}{10.5} = 0.34$$

(iii) Coefficient of mean deviation (about mode) from the table

$$= \frac{36}{10}$$

12

$$= \frac{3.6}{12} = 0.30$$

Or

$$\frac{44}{10} = \frac{4.4}{14} = 0.31$$

14

(d) Same as (cii)

3.7 Standard Deviation

This is the most important and commonly used measure of spread, especially, in statistical estimations. It is simply defined as the square root of the variance.

Thus, sample standard deviation for ungrouped data and grouped data are given as:

$$S = \sqrt{\frac{\Sigma(x - x)^2 / n}{n - 1}} \quad \text{and} \quad S = \sqrt{\frac{\Sigma(x - x)^2 / n}{n - 1}}$$

For simplicity and easy computation, the two can be reduced to-

$$S = \sqrt{\frac{\Sigma fx_i^2 - (\Sigma fx_i)^2/n}{n - 1}} \quad \text{and} \quad S = \sqrt{\frac{\Sigma fx - (\Sigma fx_i)^2/n}{n - 1}}$$

The population standard deviation for ungrouped and grouped data is-

$$\sigma = \sqrt{\frac{\Sigma (xi - \mu)^2}{N}} \quad \text{and} \quad \sigma = \sqrt{\frac{\Sigma f (xi - \mu)^2}{N}}$$

and it is simplified thus-

$$\sigma = \sqrt{\frac{\Sigma X^2_i - (\Sigma X_i)^2/N}{N}} \quad \text{and} \quad \sigma = \sqrt{\frac{X^2_i - (\Sigma X_i)^2/N}{N}}$$

Example 8

Given the sample data 1, 2, 3, 4, 5, compute the standard deviation.

Solution

$$S = \sqrt{\frac{\Sigma X^2_i - (\Sigma X_i)^2}{n}}$$

$$\Sigma X^2_i = 1^2 + 2^2 + 3^2 + 4^2 + 5^2$$

$$= 1 + 4 + 9 + 16 + 25 = 55$$

$$\therefore S = \sqrt{\frac{55 - (15)^2}{5}} = \sqrt{\frac{55 - 45}{4}} = \sqrt{\frac{10}{4}}$$

2.5

Example 9

Compute the sample standard deviation of the data below.

Table 2.3

Age	5-14	15-24	25-34	35-44	45-54	55-64
Number if cases	5	10	120	22	13	5

Age	F	Midpt (x)	Fx	X^2	Fx^2	d = x – A	$d=^d/_c$	Fd	Fd^2
5 -14	5	9.5	47.5	90.25	45.25	-30	-3	-15	45
15- 24	10	19.5	195	380.25	3802.50	-20	-2	-20	40
25 – 34	120	29.5	3540	870.25	104.430	-10	-1	-120	120
35 – 44	22	39.5	869	1560.25	34325.50	0	0	0	0
45– 54	13	49.5	643.5	2450.25	31853.25	10	1	13	13
55-64	5	59.5	297.5	3540.25	17701.25	20	2	10	20
	175		Σfx =5592.5		Σfx^2 =192563.75			Σfd= -132	Σfxd^2 =238

$$S = \sqrt{\frac{\Sigma fX_i^2 - (\Sigma fX_i)^2/n}{n-1}}$$

$$= \sqrt{\frac{192563.75 - \frac{(5592.5)^2}{175}}{175 - 1}}$$

$$= \sqrt{\frac{}{175.1}}$$

$$= 8.92.$$

OR

Using $S = \sqrt{\frac{\Sigma fd^2 - (\Sigma fd)^2}{n}}$ $= \sqrt{\frac{238 - \frac{(-132)^2}{175}}{}}$

The last formula is simplified = 8.92 than the first.

3.8 Variance

The variance is simply the standard deviation squared. Thus, the sample and population variances are given by-

$S^2 = \frac{\Sigma(xi - \bar{x})^2}{n-1}$ and

$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$ For the purpose of simplification, we use-

$$S^2 = \frac{\Sigma fd^{12} - \frac{(\Sigma fd^1)^2}{n}}{n-1}$$

Example 10

Compute the sample variance in example 19.

Solution

$$S^2 = 238 - \frac{(5592.5)^2}{175} \cdot \frac{}{175.1}$$

$$= 8.92^2$$

$$= 79.560$$

$$= 79.6$$

3.9 Coefficient of Variation

This a measure of relative dispersion for a data set, found by dividing the standard deviation by the mean and multiplying by 100 to express the coefficient as a percentage. Its sample and populations measure are given as:

$$CV = \frac{S}{\overline{X}} \times 100 \text{ or } \qquad CV = \frac{\sigma}{\mu} \times 100\%$$

Example 11

Find coefficient of variation in example 19-

$$CV = = \frac{S}{\overline{X}} \times 100\%$$

$$= \frac{8.92}{\overline{X}} \times 100\%$$

$$\overline{X} = A + \frac{\Sigma fd^1}{\Sigma f} \times W$$

$$= 39.5 + \frac{(-132)}{175} \times 10$$

$$= 39.5 - 7.54$$

$$= 31.96$$

$$C.V = \frac{8.92}{31.96} \times 100$$

$$= 27.9\%$$

Example 12

The department of community affairs is analysing the apartment rental rates in the northern, central and southern parts of Nigeria. The mean rental rates are N630, N595, and N505 in the north, central and south, respectively. The standard deviation of the rental rates are N105, N90, and N85 in the north, central and south. In which part of the country area are rental rates relatively more variable?

Solution

$$CV = \frac{S}{x} \times 10\%$$

$$CV_N = \frac{105}{630} \times 100\% = 10.67\%$$

$$CV_C = \frac{90}{595} \times 100\% = 10.51\%$$

$$CV_S = \frac{85}{505} \times 100\% = 10.68\%$$

Therefore, in a relative sense, the department should conclude that the southern part is most variable and the central is least, because 10.68% is larger than 10.51%, i.e. the standard deviation in the south is a larger proportion of the mean and standard deviation in the central area is a smaller proportion of the mean.

3.10 Relationship between Measures of Dispersion

For moderately skewed distributions, we have MD (Mean Deviation) = 4(SD) and semi-inter-quartile range:

$$\frac{(Q3 - Q1)}{2} = \frac{4}{3} SD$$

3.11 Moment

The moment of a random variable *X* is defined as the mean of powers of *X*. In which case, the r^{th} moment is the expectation of the r^{th} power of *X*.

The moment formula is given as-

$$Y^r = \frac{\Sigma Y^r}{N} \quad \_\_\_\_\_ * * *$$

Then the central moment or moment about the mean is-

$$M^r = \frac{\Sigma(x - \overline{X})^r}{N} \quad \_\_\_\_\_ *****$$

When r = 1, * * * becomes

$$Y = \frac{\Sigma Y}{N}$$

i.e. $U^1_1 = \Sigma(x)$, which is the mean of X.

When r = 1, * *** becomes

$$M^1 = \Sigma(x - x)$$
$$= \Sigma x - nx$$
$$= \Sigma x - \Sigma x, \text{ as } x = \frac{\Sigma x}{n} = nx = \Sigma x$$

When r = 2

$$M^2 = \frac{\Sigma(x - x)^2}{N},$$ which gives variance.

So that second central moment or moment about the mean is variance of *X*.

Example 13

Calculate the: (a) 2nd moment (b) 3rd moment and (c) 5th central moment of the set of data given below.

1, 3, 4, 5

Solution

(a) $Y = \Sigma^{n}_{r=1} Y^{r}$

$r = 2$

$Y^2 = \frac{1^2 + 3^2 + 4^2 + 5^2}{4}$

$= \frac{1 + 9 + 16 + 25}{4} = 12.75$

(b) $Y^3 = \frac{1^3 + 3^3 + 4^3 + 5^3}{4}$

$= \frac{1 + 27 + 64 + 125}{4} = 54.25$

(c) $M^5 = \frac{\Sigma(xi - \overline{X})}{N}$

When r = 1, $Y1 = \frac{1+ 3 + 4 + 5}{4} = x = \frac{13}{4} = 3.25$

$M5 = \frac{(1-3.25)^5 + (3-3.25) + (4 - 13.25)^2 + (5 - 3.25)^5}{4}$

$= -10.26$

3.11.1 Moment of Grouped Data

To illustrate this, you have to try your hands on the example below.

Example 14

Compute the first central moment of the data set below.

Class	**Class Frequency**
1 -5	2
6 – 10	5
11 – 15	12
16 – 20	6

Solution

Table 2.4

Class	Midpt(x)	F	Fx	F(x –x)	F(x-x)2	F(x-x)3	F(x – x)3
1 -5	3	2	6	-18.8	176.72	-1661.168	14327.86
6– 10	8	5	40	-22	96.8	425.92	1874.048
11-15	13	12	156	7.2	4.32	5.292	1.5552
16-20	18	6	108	33.6	188.16	1053.696	5900.698
		Σf=25	Σfx =310	Σf(x-x)=0	$\Sigma f=(x-x)^2$ =466	$\Sigma f=(x-x)^2$ =176.26	$\Sigma f=(x-x)^2$ =22104.1612

$$M^r = \Sigma f = \frac{(x-x)^4}{N = \Sigma f} = 0$$

$$M^2 = \Sigma f = \frac{(x-x)^2}{N} = \frac{466}{\Sigma f\ 25} = 18.64$$

$$M^3 = \Sigma f = \frac{(X-\overline{X})^3}{N}$$

$$= \frac{-176.26}{25}$$

$$= -7.0504$$

$$M^4 = \Sigma f = \frac{(x-x)^4}{N}$$

$$= \frac{22104.1612}{25}$$

$$= 884.17$$

3.12 Skewness

Skeweness measures the symmetry (or lack of symmetry) of a data set around the mean. It describes how much a given distribution of data differs from the perfectly symmetrical or normal bell- shaped curve. If the tail of the distribution is emanating from the left, it is negatively skewed. Whereas, if the tail of the distribution is emanating from the right, it is positively skewed, otherwise, the distribution is normal. The figures below described the two cases.

Figure 2.1

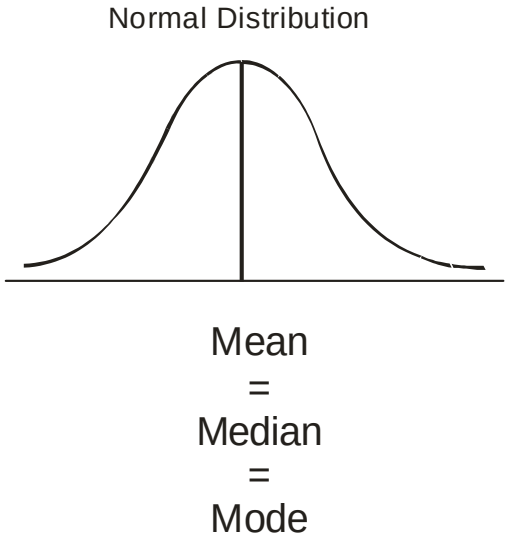


Figure 2.2

(a) Pearson's coefficient of skewness (SK)

$$= \frac{3(\text{mean} - \text{median})}{SD}$$

Some statisticians have dichotomised the above into two:

(i) Pearson's first coefficient of skewness

$$= \frac{\text{Mean} - \text{Mode}}{2}$$

(ii) Pearson's second coefficient of skewness

$$= \frac{3(\text{Mean} - \text{Median})}{SD}$$

Others are-

Quartile coefficient of skewness $= \dfrac{Q_3 + Q_1 - 2Q_2}{Q3 - Q_1}$

Percentile coefficient of skewness

$$= \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}$$

Decile coefficient of skewness

$$= \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

Moment coefficient of -

$$= \text{Kurtosis} = (a_4)\ \frac{M_4}{S_4} = \frac{M_4}{(M_2)^4}$$

$$= \sqrt{\frac{M_4}{M_2^2}}$$

Coefficients that fall between *t3* are generally considered symmetrical.

Example 15

Find the Pearson's (a) first coefficient of skewness (b) Second coefficient of skewness (c) third central moment of the set: 3, 3, 3, 4, 5,

Solution

Table 2.5

X	X – x	$(x – x)^2$	$(x – x)^3$	$(x – x)^3$
3	-1	1	1	1
3	-1	1	1	1
3	-1	1	1	1
3	-1	1	1	1
4	0	0	0	0
5	1	1	1	1
7	2	4	8	16
28 $X = {}^{28}/_{7}$		$\Sigma(x-x)^2$ = 9	$\Sigma(x-x)^3$ = 13	$\Sigma (x-x)^4$ = 21

(a) Pearson's first coefficient of skewness

$$= \frac{\text{Mean – Mode}}{\text{S.D}}$$

Mode = 3

Mean = 4

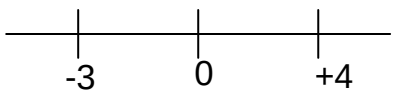
$$\text{S.D} = \sqrt{\frac{\Sigma(x-x)^2}{\text{P1st Coeff(8k)}}} = \sqrt{\frac{9}{7}} \quad \frac{4-3}{1.136} = \sqrt{1.29} = 0.880$$

Since, it is between (±3), it is symmetrical, i.e. normal.

(b) $\text{P2nd (sk)} = \frac{3(\text{mean - median})}{\text{S.D}}$

Median = (n + 1)th observations. $({}^{8}/_{2})$th obs = 4^{th} obs = 3

$$\text{P2nd (SK)} = \frac{3(4-3)}{\text{S.D}} = \frac{3}{1.136} = 2.64$$



T
It is a normal curve.

(c) The third central moment is the moment coefficient of skewness.

$$M3\ (SK) = \frac{M_3}{\sqrt{M^3_2}}$$

$$\text{Where } M_3 = \frac{\Sigma(x - x)^3}{N}$$

$$M_2 = \left(\sqrt{\frac{\Sigma(x - x)^2}{N}}\right)$$

$$= \frac{\sqrt[13]{/_7 M^3_2}}{\left(\sqrt{\frac{9}{7}}\right)} = \frac{\left(\sqrt{\frac{\Sigma(x - x)^2}{1.857N}}\right)^3}{1.136^3} = \frac{1.857}{1.466}$$

$$= 1.267$$

It is symmetrical, since 1.267 falls between ±3.

3.13 Kurtosis

This is the degree of the peak or flatness in the centre of the symmetric distributions. These forms of departures may be detected by the coefficient of kurtosis for the sample-

$$\text{Kcoeff} = \frac{\Sigma(x - x)^4}{\left(\sqrt{S^2}\right)^4}$$

$$= \frac{\text{4TH Central moment}}{(\text{standard deviation})^4}$$

A relatively high peak, with relative wide tails is called *leptokurtic*, meaning a "narrow humped". A flat-topped and ,relatively, tin tail-meaning "broad humped" is called platy-kurtic; and a normal distribution which is either high peaked nor flat-topped. It is a sort of normal distribution setting where other values are evenly distributed around the mean. This is called *mesokurtic*. We can also use quartile and percentile to measure the degree of peakness of a distribution; thus, moment coefficient of kurtosis.

$$\text{Kcoff} = \frac{\text{Semi-inter quartile range}}{\text{Inter-percentile range}}$$

$$= \frac{(Q_3 - Q_2)/2}{P_{90} - P_{10}}$$

Looking at the peakness of a distribution, we shall draw the axiom that, for a normal distribution the moment coefficient of kurtosis is $a_4 = 3$. Thus, if moment coefficient is greater than 3, i.e. $a_4 > 3$ it is *leptokurtic*; but if $a_4 < 3$, it is *platykurtic*.

Example 16

Compute the 4th central coefficient of kurtosis.

$$\text{Kedeff.} = \frac{M^4}{\left(\sqrt{S}\right)^4}$$

$$= \frac{{}^{21}/_{7}}{1.1364} = 3 = 1.80$$

Since, $a_4 < 3$, it is *platykutic*.

$$Q_3 = 3\frac{(n+1)}{4}\text{th obs} = 6^{th}\text{ obs} = 5$$

$$Q_1 = \frac{(n+1)}{4}\text{th obs} = 3\text{th obs} = 3.$$

$$P_{90} = \frac{90}{100}(n+1)\text{t hobs} = 2^{nd}\text{ 0bs} = 5.6$$

$$= 7 + 0.2\,(0-7)$$
$$= 7 + 0.2(-7)$$
$$= 7 - 1.4$$
$$= 5.6$$

$$P_{90} = \frac{10}{100}(n+1)\text{t hobs} = 0.8^{th}\text{ obs}$$

$$= 0 + 0.8\,(3-0)$$
$$= 0 + 2.4$$
$$= 2.4$$

$$\text{Kcoeff} = \frac{5-3}{5.6-2.4}$$

$$= {}^{1}/_{3.2} = 0.3125$$

$a_4 < 3$, *it is plalykurtic*

SELF-ASSESSMENT EXERCISE

In a survey of small breweries in the south-east, 10 such firms report the following numbers of employees: 15,14,12,19,13,14,15,18,13,19. Compute: (a) first moment (b) second moment (c) third moment (d) first central moment (e) second central moment (f) third central (g) fourth central moment (h) Pearson's coefficient of skewness (i) moment coefficient of kurtosis.

4.0 CONCLUSION

This unit has enabled you to discover what to do whenever you want to measure the spread of a quantitative data. By now you would have come to realise that the computation is easy and straightforward. Note that standard deviation is the most important of all measures of variability.

5.0 SUMMARY

In this unit, you have learnt how to define and compute some measures of variability such as coefficient of variation, population co- efficient of variation and so on.

6.0 TUTOR-MARKED ASSIGNMENT

i. Two regionally distributed, general-interest magazines solicit advertising from the same clientele. The advertising rates and the total number of subscribers are the same for both magazines. The following table shows the age distribution of subscribers of the magazines.

Age (Yrs)	**Magazine A**	**Magazine B**
10 – 19	1	1
20 – 29	5	2
30 – 39	6	2
40 – 49	4	3
50 – 59	3	5
60 – 69	1	4
70 – 79	1	3

Compute: (i) the coefficient of variation and give your comment (ii) P_{90} of magazine A (iii) quartile deviation of magazine B (iv) D_5 and D_1 of magazine B. (v) 99^{th} percentile of magazine A and B.

ii. Calabar Rovers team scored the following number of points in each of their last ten matches respectively-

18,3,21,159,84,27,10,442,6,15. Compute the following descriptive statistics of variability for the set of points; (a) range (b) upper and lower quartiles (c) inter-quartile (d) semi-inter-quartile range (e) upper decile (f) standard deviation (g) variance (h) coefficient of variance (i) quartile coefficient of dispersion.

7.0 REFERENCES/FURTHER READING

Brian, P.M. & Philip, M.N. (2006). *Applied Statistics for Public Policy*. New Delhi: Prentice-Hall.

James, T. & Benon, P.G. (1988). *Statistics for Business and Economics*. (4th ed.). United States: Dellen Publishing Company Inc.

Okafor, R. (2004). *Statistical Methods Plus Non Parametric Techniques*. JAS Publishers.

UNIT 3 PROBABILITY

CONTENTS

1.0 INTRODUCTION

So far, you have learnt about measures of location and variability. This prior knowledge will enhance your understanding of probability. Simply put, all humans are faced with uncertainty about life; thus, probability theory is a basic component of decision-making, under uncertainty. From the sciences, arts to the humanities we are confronted with decision making under uncertainty. The objective of this unit then, is to present to you the basic concept of probability needed for an understanding of statistical inference. We shall introduce the subject by considering probabilities that are based on observed data.

2.0 OBJECTIVES

At the end of the unit, you should be able to:

- highlight the role of probability in drawing inferences, in the face of uncertainty
- assign probabilities using classical, relative, Bayesian and axiomatic models
- describe permutation and combination.

3.0 MAIN CONTENT

3.1 Definition of Terms

Here, let us look at a number of terms

a. Random equipment

This is an experiment which the outcome is unpredictable- for example, tossing a coin, rolling a die, lottery card game etc.

b. Sample space

This is the set of all possible outcome of a random experiment. This sample space is denoted by Ω, and a generic element of the sample space is denoted by ω. For example, driving to work, a commuter passes through a sequence of three intersections with traffic lights. At each light, he either stops, (s) or continues (c). The sample space is the set of all possible outcomes.

Ω = {CCC, CCS, CSS, CSC, SSS, SSC, SCC, SCS}
Where, *CSC*- for example, denotes the outcome of the movement of the commuter- through the first light, stopping at the second light and continuing through the third light.

c. Event

Here, we are interested in particular subset of Ω, which in probability language are called events. We denote events by capital alphabets A, B, C,… X, Y, Z. In the example above, the event that the commuter stops at the first light is the subset of Ω denoted as follows-

A = {SSS, SSC, SCC, SCS}.

d. Intersection of two events

A set *M* is said to be intersection of two sets *A* and *B* if the elements of *M* are the elements common to both *A* and *B*. *M* is denoted as $M = A \cap B$. For example, if *A* is the event that the commuter stops at the first light (listed above) and if *B* is the event that he stops at the third light, therefore-

B = {CCS, CSS, SSS, SCS)
M = $A \cap B$ = {SSS, SCS)

e. **The union of two sets**

A and *B*, is the event *M* that either *A* occurs or *B* occurs or both occur- *M = A U B*. For example, if *A* is the event that the commuter stops at the first light (see above) and *B* is the event that he stops at the third light (see above), therefore:
M = {SSS, SSC, SCC, SCS, CCS, CSS}.

f. Impossible events

An impossible event is an event that will never happen. It is represented by φ. For example, $T = \{x: 0 < 1\ x < ,\ x \Sigma N\}$ is an impossible event, where *N* is the set of natural numbers.

g. Sure events or certain event

An event that must happen is called sure or certain event. For example, in tossing a die-
$T = \{x : 1 \leq x \leq 6\}$ is a sure event.

h. Mutually exclusive events

Two events- *P* and *Q* are called mutually exclusive event or disjointed events if the occurrence of one of them precludes (prevents) the occurrence of the other. That is, $P \cap Q = \varphi$. For example, in rolling a die, where *P* and *Q* are numbers on the upturned face of the die.
Let P = {x : P is even}
Q = {x : Q is odd)

Then $P \cap Q = \varphi$, which- when interpreted, *P* and *Q* are mutually exclusive.

i. Complementary events

The complement of an event- A^c, is the event that *A* does not occur, and thus consist of all those elements in the sample space that are not in *A*.
That is, if $A \cup A^c =$ ∩ *or* $A^c = \Omega - A$.
For example, the complement of the event that the commuter stops at the first light is the event that he continues at the first light.
:. A^c = {CCC, CCS, CSC}

j. Event space (∝)

The set of all possible subset of a sample space is called the event space, it is most often called the power set in elementary mathematics. For example, 4 ∩ = { a, b)

$\mathcal{P}(S)$ = {a}, {b}, {a,b}, {φ}}
= {(a}, {b} ($\cap$){φ}}

k. Independent event

Two events *P* and *Q* are said to be independent if the occurrence (or non-occurrence) or *P* has no effect on the occurrence (or non-occurrence) of *Q*. For example, toss a coin twice, and each time, observe the face that shows up. The events-
P = {heads in first toss}
Q = {heads in second toss}

3.2 Definition of Probability

The earliest definition of probability is the classical or objective definition, which is based on repeated trial of tests and the assumption that various outcome of a trial are equally, likely and mutually exhaustive events. For example, if there are *n* exhaustive, equally, likely and mutually exclusive cases- *"y"* of which are favourable to the occurrence of event T, then *Pr (T)* $= {}^{y}/_{n}$
For example, if *S* = {1, 2, 3, 4, 5, 6}. Find the probability of getting a number less than 6. These are six possible cases, out of these 6 cases, only 5 numbers are less than 5 i.e: T = { 1, 2, 3, 4, 5}.
Therefore, P_r (T) = ${}^{y}/_{n}$ = ${}^{5}/_{6}$

3.3 Frequentist or Relative Probability

When the number of trails is carried out in an infinite number of repeated tests, we can define the relative frequency *f* the event (P) in terms of

$$Rf(P) = \frac{\text{Number of trials in which P occurs}}{\text{Total number of trails}}$$

This ratio *Rf (P)* is called the empirical probability. This definition assumes that probability of an event has a limiting value on the long run.

$$Pr(P) = \lim_{n \to \infty} = \frac{y}{n}.$$

We, therefore, conclude that the definitions by the classical and relative frequency schools of thought are based on objectivity.

3.4 Bayesian or Subjective Probability

According to the Bayesian model, probability is not based on repeated trials of tests. For Bayesian then, probability is a model for quantifying the strength or personal options. Baye's rule describes how personal opinion evolves with experience. Suppose that the prior probability of *A* is *P (A)*. On observation of an event *P*, the opinion about *P* changes to *Pr (A/P)* according to Baye's rule-

$$\Pr(A/P) = \frac{\Pr\{P/A\}\ \Pr\{A\}}{\Pr(P)}$$

You will learn more about this, as we proceed in this course.

3.5 Axiomatic Probability

An axiom is a rule or principle that is, generally, believed to be true. It does not need to be proved or debunked. In the axiomatic approach, we develop an idealised model from which we can predict the probability of the occurrence of various events. This model establishes abstract relationship between the events of a random experiment, and as such, could be used in calculating their probabilities. For example-

Pr (A or B) = Pr (A) + Pr (B) - P (A n B) – non mutually exclusive event is an axiomatic approach.

Example 1

(1) A die is rolled once; what are the probabilities of getting:

i. an even number
ii. an odd number
iii. a prime number
iv. an event prime number
v. an odd prime number.

Solution

Let S = { 1, 2, 3, 4, 5, 6}
Let P = event that even number occurs
Q = event that odd number occurs
M = event that prime number occurs
L = event that even prime number occurs
F = event that odd prime number occurs

(i) n (S) = 6
P {2, 4, 6}
n (P) = 3
$Pr(P) = \frac{n(P)}{n(S)} = \frac{3}{6} = \frac{1}{2}$

(ii) Q {1, 3, 5}
n (Q) = 3
$Pr(Q) = \frac{n(Q)}{n(S)} = \frac{3}{6} = \frac{1}{2}$

(iii) M {2, 3, 5}
n (M) = 3
$Pr(M) = \frac{n(M)}{n(S)} = \frac{3}{6} = \frac{1}{2}$

(iv) L {2}
n (L) = 1
$Pr(L) = \frac{n(L)}{n(S)} = \frac{1}{6} = {}^{1}/_{6}$

(v) F {3, 5}
n (F) = 2
$Pr(F) = \frac{n(F)}{n(F)} = \frac{2}{6} = {}^{1}/_{3}$

Example 2

If four fair coins are tossed, find the probability of getting:

i. at least, 3 heads
ii. at most, one head
iii. exactly, four heads

Solution

Let the sample space be *S* and at least 3 heads, at most one head and exactly four heads be *M*, *N* and *O* respectively.

Then-
S = {(H H H H), (HHHT), (HHTH), (HHTT), (HTHT)}
{(HTTH), (HT,TT), (THHH),(THHT)(THTH)
{(THTT), (TTHH), (TTHT),(TTTH),(TTTT)}
n (S) = 16
M = {(HHHH), (HHHT), (HTHH),(HHTH),(THHH)}
n (M) = 5
N = {(HTT),(THTT),(TTHT),(TTTH)}
N (N) = 4
O = {HHH}
n (O) = 1

(1) $\Pr(M) = \frac{n(M)}{n(S)} = \frac{5}{6}$.

(2) $\Pr(N) = \frac{n(N)}{n(S)} = \frac{4}{16} = \frac{1}{4}$.

(3) $\Pr(O) = \frac{n(O)}{n(S)} = \frac{1}{16}$.

3.6 Basic Laws of Probability

You are to note the following.

(a) $0 \leq P(E) \leq 1$

(b) $\Pr(\cap) = 1$, $\Pr(\varphi) = 0$

(c) For any E $\cap$, and its complement E^c, we have:
$\Pr(E) + \Pr(E^c) = 1$

(d) For any events E_1 and E_2
$\Pr(E_1, U, E_2) = P(E_1) + \Pr(E_2) – (E_2) – \Pr(E_1 \text{ n } E_2)$

3.7 Conditional Probability

The conditional probability of a given *B* is denoted by Pr (A/B). Similarly, the conditional probability of *B* given *A* is represented by *Pr(B/A)*.

That is-

$\Pr(A/B) = \frac{\Pr(A \text{ n } B)}{\Pr(B)}$ provided $\Pr(B) \neq 0$

Example 3

If *A* and *B* are independent events with Pr(A) = 0.05 and Pr(B) = 0.65
Find- Pr (A/B)

Solution- $\Pr(A/B) = \frac{\Pr(A \text{ n } B)}{\Pr(B)} = \frac{\Pr(A) \times \Pr(B)}{\Pr(B)} = \Pr(A) =$ 0.05 Pr(B)=0.65

Pr(A/B) = Pr(A) x Pr(B)

= 0.05 x 0.65 = 0.325

Example 4

The work force of a firm-running to about 200 people comprises four groups, namely- male manual workers, male supervisors, female manual workers and female supervisors, as follows:

	Male (M)	**Female (F)**
Supervisors (S)	20	50
Manual workers (W)	100	30

If a supervisor is selected at random, find the conditional probabilities.

(a) Pr (M/S) (b) Pr(F/S) (c) Pr (S/F)

Solution

	Male (M)	**Female (F)**	**Total**
Supervisors(S)	20	50	70
Manual workers(W)	100	30	130
	120	80	200

(a) $\Pr(M/S) = \dfrac{\Pr(M \cap S)}{\Pr(S)} = \dfrac{20}{200}$

$\Pr(M \cap S) = \dfrac{70}{200}$

Therefore- $\Pr(M/S) = \dfrac{\frac{20}{200}}{\frac{70}{200}} = \dfrac{20}{200} \times \dfrac{200}{7} = \dfrac{2}{7}$

(b) $\Pr(F/S) = \dfrac{\Pr(F \cap S)}{\Pr(S)} = \dfrac{50}{200}$

$\Pr(S) = \dfrac{70}{200}$

Therefore, $\Pr(M/S) = \dfrac{\frac{50}{200}}{\frac{70}{200}} = \dfrac{50}{70} = \dfrac{8}{7}$

(c) $\Pr(S/F) = \dfrac{\Pr(S \cap F)}{\Pr(f)} = \dfrac{\frac{50}{200}}{\frac{80}{200}} = \dfrac{50}{80} = \dfrac{5}{8}$

3.8 Multiplication Law of Probability

From our earlier definition, we have that-

$$\Pr(A/B) = \frac{\Pr(A \cap B)}{P(B)}$$

By simple cross-multiplication, we have-

$$\Pr(A \cap B) = \Pr(A/B) \times \Pr(B)$$

Similarly,

$$\Pr(B/A) = \frac{\Pr(B \cap A)}{\Pr(A)}$$

$$\Pr(B \cap A) = \Pr(B/A) \times \Pr(A)$$

Example 5

A box contains 6 red balls and 4 white balls. A ball is chosen at random from the box, and then a second ball is drawn at random from the remaining balls in the box. Find the probability that:

(i) both balls are white
(ii) both balls are red
(iii) the first ball is white and the second is red
(iv) the second is red
(v) the second is white.

Solution

Let *A* be the event the first ball is red Let
B be the event the first ball is white Let
A_1 be the event the second ball is red
Let B_1 be the event the second ball is white

(i) Pr (Both ball are white) $= \Pr(B \cap B_1) = \Pr(B_1/B) \times \Pr(B_1)$

$\Pr(B) = {}^4/_{10}$

$\Pr(B_2/B) = {}^3/_9 = {}^1/_3$

Thus- $\Pr(B \cap B_1) = {}^4/_{10} \times {}^1/_3$

$= {}^4/_{30}$

(ii) Pr (Both ball are red) $= \Pr(A \cap A_1) = \Pr(A_1/A) \Pr(A)$

$\Pr(A) = {}^6/_{10}$

$\Pr(A_1/A) = {}^5/_9$

Thus- $\Pr(A \cap A_1) = {}^5/_9 \times {}^6/_{10}$

$= {}^1/_3$

(iii) Pr (1st ball is white and second is red)

$= B \cap A_1$

$= Pr(A_1/B)\ Pr(B)$

$Pr(B) = {}^4/_{10}$

$Pr(A_1/B) = {}^6/_9 = {}^2/_3$

Thus, $B \cap A_1 = {}^4/_{10} \times {}^2/_3 = {}^4/_{15}$

(iv) Pr (the second is red) $= Pr(A_1)$

$= Pr(A_1 \cap B) = Pr(A_1 \cap B)$

$= {}^1/_3 + {}^4/_{15}$

$= \frac{5 + 4}{15} = \frac{9}{15} = \frac{3}{5}$

(v) $Pr(B_1) = Pr(B_1 \cap B) + Pr(B_1 \cap A)$

$= {}^4/_{30} + {}^4/_{15} = \frac{4 + 8}{30} = \frac{12}{30} = \frac{2}{5}$

3.9 Marginal Probability

This is when we ignore one or more criteria of classification in computing a probability. For instance, in example 4 above, if our interest is to compute the probability of a person selected at random, it will be a supervisor. This is marginal probability, because, interest is centred on a probability associated with a marginal total, and we disregard another criterion for classification.

i.e. $Pr(S) = \frac{70}{200} = \frac{7}{20}$

In other words-

$Pr(S) = Pr(S \cap M) + Pr(S \cap F)$

$= \frac{20}{200} + \frac{50}{200} = \frac{70}{200}$

$Pr(W) = Pr(W \cap M) + Pr(W \cap F)$

$= \frac{100}{200} + \frac{30}{200} = \frac{130}{200} = \frac{13}{20}$

3.10 Additional Rule

Given two events *A* and *B,* the probability that event *A* or event *B* or both occur is equal to the probability that event *A* occurs, plus the probability that event *B* occurs, minus the probability that both events occur.

Axiomatically, if we are to use example 4 above, it is written as-

$$\Pr(S \cup M) = \Pr(S) + \Pr(M) - \Pr(S \cap M)$$
$$= \frac{70}{200} + \frac{120}{200} - \frac{20}{200}$$
$$= \frac{210 - 20}{200} = \frac{90}{200} = \frac{9}{20}$$

3.11 Joint Probability

By joint probability, we mean two events that are independent. That is-
$(\Pr(A_1 \cap A_2) = \Pr(A_1) \times \Pr(A_2)$

3.12 Bayes Theorem

Given $X_1, X_2, \ldots, X_n$, mutually exclusive events whose union is the universe, and let *A* be an arbitrary event in the universe, such that P (A) ≠ 0, then

$$\Pr(X/A) = \frac{\Pr(A/x_j)\ \Pr(A)}{\sum_{i=1}^{n} \Pr(A/X_i)\ \Pr(A)}$$

Where i = 1, 2, …n, j = 1, 2, …n

The mathematical formular above is called the formular for the probability of cause, since it enable us to find the probability of a particular B_j or "cause" by which *A* may have been brought about it. It is sometimes written in another form as follows $\Pr(B_j/A) = \frac{\Pr(A \cap Bj)}{\Pr(A)}$

Where *Pr(A)* is the marginal probability of *A* which is defined from the law of total probability.

$$\text{Thus } P(A) = \sum_{i=1}^{n} \Pr(A \cap B_i)$$
$$= \Pr(A \cap B) + \Pr(A \cap B_2)$$
$$= + \ldots + \Pr(A \cap B_n)$$
$$= \frac{\Pr(B_i/A) = \Pr(A \cap B_i)}{\Pr(A \cap B_i) + \Pr(A \cap B) + \ldots + \Pr(A \cap B_n)}$$

Example 6

Of the persons employed by a large manufacturing firm, 45% come from one area of the city- area *A*, 30% from second area- *B*, and 25% from a third area- *C*. Within one year, 30% of the employees from area *A*, 20% from area *B* and 10% from area *C* left the firm. When a record is picked at random from the firm's files, it is found that the person under consideration left the firm within one year. What is the probability that the person was from area A, area B and area C?

Let *M* designate the event that a record is picked at random. And let *A, B* and *C* represent the event that the persons employed came from Area *A, B* and *C,* respectively. We are to compute:

Pr (A / M), Pr (B/ M) and Pr (C / M)

Pr (A)	=	45%	=	0.45
Pr (B)	=	30%	=	0.3
Pr (C)	=	25%	=	0.25
Pr (M / A)	=	30%	=	0.3
Pr (M / B)	=	20%	=	0.2
Pr (M / C)	=	10%	=	0.1

These probabilities are called the likelihood. From this probability, we can calculate the three joint probabilities.

Pr (M n N)	=	Pr (M / A)	Pr (A) = 0.3 x 0.45	=	0135
Pr (M n B)	=	Pr (M / B)	Pr (B) = 0.2 x 0.3	=	0.06
Pr (M n C)	=	Pr (M / C)	Pr (C) = 0.1 x 0.25	=	0.025

Summary or calculations illustrating the use of Baye's theorem.

Event	***Prior Probability***	***Likelihood Probability***	***Joint Probability***	***Postering Probability***
A	0.45	0.3	0.135	0.61
B	0.3	0.2	0.06	0.27
C	0.25	0.1	0.025	0.11
Total	1.00	0.22	1.09	

$$\Pr(A / M) = \frac{\Pr(M / A)\ P(A)}{\Pr(M \cap A) + \Pr(M \cap B) + \Pr(M \cap C)} = \frac{0.135}{0.22} = 0.61$$

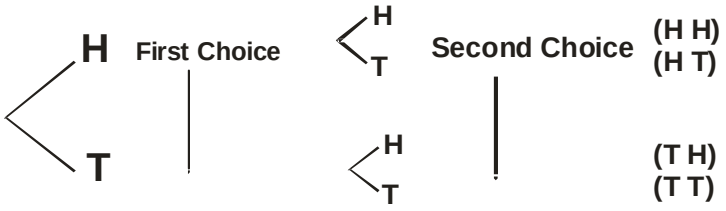
$$\Pr(B / M) = \frac{\Pr(M / B)\ \Pr(B)}{\Pr(M \cap B) + \Pr(M \cap N) + \Pr(M \cap C)} = \frac{0.06}{0.22} = 0.27 .$$

$$\Pr(C / M) = \frac{\Pr(C \cap M)}{\Pr(M)} = \frac{0.025}{0.22} = 0.1$$

3.13 Tree Diagram

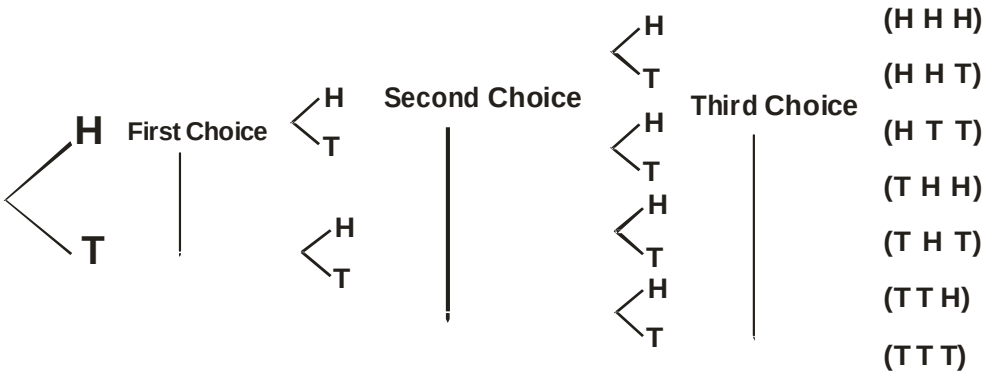
A tree diagram shows all the possible combination or outcome in a random experiment. For example, if a coin is thrown twice, we can use a tree diagram to display the entire possible outcome. It is also used in business decision analysis.

Thus-



Ω = ({H, T), (H, T), (T, H) (T, T)

If it is thrown thrice, we shall have the following possible outcomes



Ω {(H H H), (H H T), (H T H), (H T T) (H T T), (T H H), (T H T), (T T H), (T T T)}

3.14 Permutation and Combination

Permutations refer to the number of ways in which a set or objects can be arranged in order (the order being crucial); while combinations refer to number of ways in which a set or object can be arranged, regardless of order. Precisely, it is selection of objects where order is not very important.

The mathematical expression for permutations and combinations involve some new notations namely *n* !-which denotes the products or n(n – 1) (n - 2) (n – 3) … (2) (1).

Thus, 3! = 3 x 2 x 1 = 6

0! = 1 and
1! = !

You can derive factional from your calculator, it is seen as n! or x!, just press a particular number, then press 2ndf, (n! or x!) on the number you intend to factor out e.g. 3 2ndf (x! or n!) gives 6.

In general, permutation is written as $^{n}Pr = \frac{n!}{(n-1)!}$

Combination is written as $^{n}C_r = npr = \frac{n!}{r^1(n-1)!}$

Also note that-

$$^{n}C_r = {}^{n}C - \text{rand } {}^{n}C_r + {}^{n}C_{n-1} = {}^{n+1}Cr$$

$$^{n}Pr = \frac{n(n-1)(n-2)\ldots.(n-r+1)(n-r)(n-r-1)\ldots.1}{(n-r)(n-r-1)\ldots.1}$$

$$= n(n-1)(n-2)\ldots(n-r+1)$$

$$^{n}C_r = \frac{n(n-1)(n-2)\ldots.(n-r+1)(n-r)(n-r-1)\ldots.1}{(n-r)(n-r-1)\ldots.1}$$

$$= \frac{n(n-1)(n-2)\ldots n-r+1}{r!}$$

$$= \frac{n(n-1)(n-2)\ldots n-r+1}{r(r-1)(cr-2)\ldots.1}$$

Example 7

Evaluate (i) $8p_2$ (ii) $_{b}P_4$

Solution

$$8P_2 = \frac{8!}{(8-2)!} = \frac{8!}{6} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{6 \times 5 \times 4 \times 3 \times 2 \times 1}$$

$$= 56$$

$$10P_4 = \frac{10!}{(10-4)!} = \frac{10!}{6} = \frac{10 \times 9 \times 8 \times 7 \times 6}{6!}$$

$$= 56$$

Example 8

A supervisor has 7 workers available from which to form a 4 – member production team. How many different teams are possible?

Solution

This is combination: (7_4) i.e. chose team from 7 4 members - production workers.

$(7_4) = 7_{c4}$
$= 7!$
$(7 - 4)!\ 3!$
$= \frac{7!}{3!\ 3!} = \frac{7 \times 6 \times 5 \times 4 \times 3!}{3! \times 3! \times 2}$
$= 7 \times 5 \times 4$
$= 35 \times 4$
$= 140$ ways

Again, if nPn, n_2 …., n_k equals the number of distinguishable sequence that can be formed, *n* objects take *n* at a time, when n_1 are of one type, n_2 are of a second type … and n_k are of *kth* type, $n = n_1 + n_2 + n_k$. We may generalise the above result as follows-
$n! = (np_n, n_2 \ldots nk)\ n_1!, n_2!, \ldots n_k!$

If we solve for ${}^{n}P_{n1, n2 \ldots nk} = \frac{n!}{n!\ n_2!\ n_k!}$.

4.0 CONCLUSION

By now, your understanding of the concept of probability should have been enhanced. Decision- making is paramount in life, in business and research work. By applying the principles of probability, quality decisions can be easily arrived at. Your lives are full of choices and better alternative choices indicate a better chance for improvement and optimum performance. Applying the mathematics of probabilities you learned in this unit will make a difference. You are going to now study probability distributions in the next unit.

5.0 SUMMARY

You have learnt the following key facts in this unit.

- Addition rule for mutually exclusive events- if *A* and *B* are two mutually exclusive events, then the probability of obtaining either *A* and *B* is equal to the probability of obtaining *A* plus the probability of obtaining *B*.
  Pr (A or B) = Pr (A) + Pr (B)
- Addition rule for non- mutually exclusive events- if *A* and *B* are non- mutually exclusive events, then we must subtract the probability of the joint occurrence of *A* and *B* from the sum of their probabilities.
  Pr (A or B) = Pr (A) + Pr (B) - Pr (A ∩ B)
- Bayes's rule for conditional probabilities. A general method for revising prior probabilities in the light of new information, to provide posterior probabilities.

  Two events: $Pr(B/A) = \frac{Pr(A \text{ and } B)}{Pr(A)}$

  General form: $Pr(B/A) = \frac{Pr(A/Bj)\,Pr(Bj)}{\sum_{j=1}^{n} Pr(A/Bj)\,Pr(Bj)}$

6.0 TUTOR-MARKED ASSIGNMENT

i. A residential sub-division developer has a house styles to build on 11 adjacent lots. How many distinguishable arrangements are possible if the developer decides to build 2 houses of each style?

ii. An investment firm is interested in the following-
A = {Common stock in *NYZ* corporation gains 10% next year} B = {Gross national product gains 10% next year}. The firm has assigned the following probability distribution on the bases of available information-
Pr(A / B) = 0.8 (P (B) 0.3
Compute Pr (A n B)

iii. In a large sub-urban community, 30% of the large household use Brand A toothpaste, 27% use Brand B, 25% use Brand C, and 18% use brand D. In the four groups of households, the proportion of resident who learned about the brand they use through television advertising are as follows= Brand A, 0.10, Brand B, 0.05, Brand C, 0.02 and Brand D, 0.05; In a household selected at random from community, it is found that residents learned about the toothpaste through television advertising:

i. what is the probability that the brand of toothpaste used in the household is (a) A (b) B (c) C (d) D?
ii. Use diagram to show the prior probability, conditional probabilities, and joint probability.

7.0 REFERENCES/FURTHER READING

Brian, P.M. & Philip, M.N. (2006). *Applied Statistics for Public Policy*. New Delhi: Prentice-Hall.

James, T. & Benon, P.G. (1988). *Statistics for Business and Economics*. (4th ed.). United States: Dellen Publishing Company Inc.

Kasumu, R. B. (2002). *Introduction to Probability Theory*. *A First Course*. JAB Publishers.

UNIT 4 PROBABILITY DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

You have learnt about the concept of probability earlier on. However, in this unit we shall discuss probability distributions under two categories- probability distribution of discrete random variables and probability distribution of continuous random variables.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- distinguish discrete and continuous random variables
- construct probability distributions from raw data
- compute means and variances of probability distribution.

3.0 MAIN CONTENT

3.1 Probability Distribution of Discrete Random Variable

The probability distribution of a discrete random variable is a table, graph, formula, or other device used to specify all possible values of the discrete random variable, along with their respective probabilities. To illustrate this, suppose we roll a dice four times and we are interested in the number of heads that such show up. The sample space becomes:

Ω = {(H H H H), (HHHT), (HHTH), (HHTT), (HTHT)}
{(HTTH), (HT,TT), (THHH),(THHT)(THTH)
{(THTT), (TTHH),(TTHT),(TTTH),(TTTT)}

Which consist of 16 elements. You will discover that the number of heads assume different value as shown in the table below.

Table 4.1: Frequency Table

HEAD	0	1	2	3	4	5
HHHH					4	
HHHT				3		
HHTH				3		
HHT			2			
HTHH				3		
HTHT			2			
HTTH			2			
HTTT		1				
THHH				3		
THHT			2			
THTH			2			
THTT		1				
TTHH			2			
TTHH		1				
TTTH		1				
TTTT	1					
TOTAL	1	4	6	4	1	16

Table 4.2
The Probability Distribution Table

X	0	1	2	3	4
F(x)	1	4	6	4	1
P(x)	$^1/_{16}$	$^4/_{16}$	$^4/_{16}$	$^4/_{16}$	$^1/_{16}$

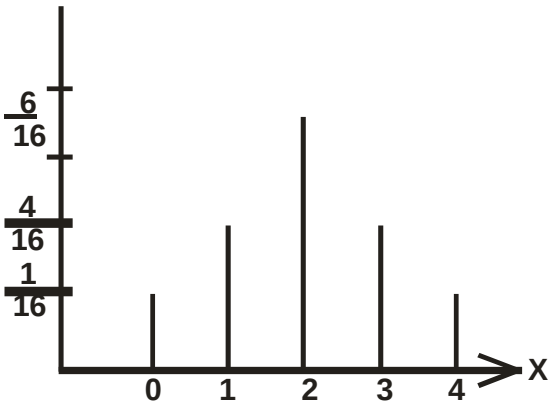


Fig. 4.1*:* Probability Distribution Graphic Form

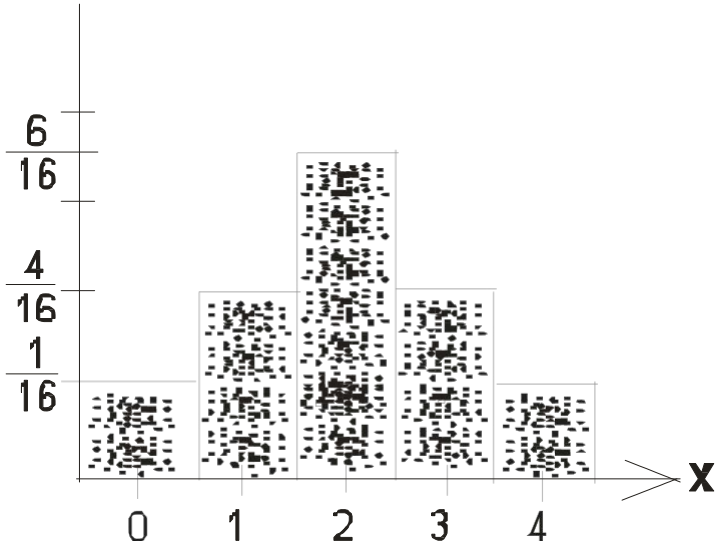


Fig. 4.2

The figures above lead to another concept- a random variable. A random variable *x* on a sample space is a function which assigns to each element ώέ one and only one real number ώέ- the sample space . In our example above, we denote the probability of the random variable *x* by the symbol *P(x)*. Then, for this example, P (0) = $^1/_{16}$, P (i) = $^4/_{16}$, P (2) = $^6/_{16}$, P (3) = $^4/_{16}$ and P (4) = $^1/_{16}$.

Again, the table above form some essential properties of the probability distribution of a discrete random variable which may be expressed as follows-

Given a discrete random variable *x* that can assume only the *K* different values X_1, X_2, X_3 ….., X_K, the probability distribution of *X* must satisfy the following two conditions:

(i) $0 < P(X = X_1) \leq 1$
(ii) $\sum P(x = x_i) = 1$, where I = 1, 2, …, K

Example 1

The random variable *X* has the following discrete probability distribution.

Table 4.3

X	0	1	2	3	4
P (X)	0.10	K	0.25	0.25	0.30

Find (a) K (b) Pr($x \leq 3$) (c) Pr ($2 \leq x \leq 4$)

Solution

(a) Since-$\sum$ Pr (X = Xi) = 1
0.1 + K + 0.25 + 0.25 + 0.30 = 1
K = 0.90 = 1
K = 1 – 0.9
= 0.10

(b) Pr($x \leq 3$) = Pr (X = 0) + Pr (X = 1) + Pr (x = 2) + Pr (X = 3)
= 0.10 + 0.10 + 0.25 + 0.25
= 0.70

(c) Pr($2 \leq x \leq 4$) = Pr (x = 2) + Pr (x = 3) + Pr (x = 4)
= 0.25 + 0.25 + 0.30
= 0.80

3.2 The Mean and Variance of Discrete Probability Distribution

The mean of a probability distribution is the expected value of the random variable that has the special distribution. The expected value of a distribution random variable *x* is merely the arithmetic mean and therefore may be labelled μ.
Similarly, we define the variance as
$E (X – \mu)^2$

$\sum (X – \mu)^2 x P (X = X)$ = $\sqrt{\ }^2$

$\sum X^2 – (\sum x)^2 = \sqrt{\ }^2$

Example 2

Consider the following probability distribution for the random variable x

Table 4.3

X	1	2	3	4	5
P (X)	0.05	0.30	0.35	0.20	0.10

Find: (a) μ (b) σ^2 (c) σ

Solution

$\mu = \sum xPr\ (X = X_i)$
$= 1 \times 0.05 + 2 \times 0.30 + 3 \times 0.35 + 4 \times 0.20 + 5 \times 0.10$
$= 0.05 + 0.60 + 1.50 + 0.80 + 0.50$
$\mu = 3$

(b) $\sigma^2 = \sum (X^2) - \sum (X)^2$
$\sum (X^2) = x^2_1\ Pr = (X = X_i)$
$= 1^2 \times 0.05 + 2^2 \times 0.30 + 3^2 \times 0.35 + 4^2 \times 0.20 + 5^2 \times 0.10$
$= 0.05 + 4 \times 0.30 + 9 \times 0.35 + 16 \times 0.02 + 25 \times 0.10$
$= 0.05 + 1.2 + 3.15 + 3.2 + 2.5$
$= 10.1$

(b) $\sigma^2 = \sum (X^2) - \sum (X)^2$
$= 10.1 - (3)^2$
$= 10.1 - 9$
$= 0.1$

(c) $\sigma = \sqrt{0.1}$
$= 3.18$

3.3 Probability Distribution of a Continuous Random Variable

A random variable that can assume values corresponding to any of the points contained in one or more interval is called a continuous random variable. Similarly, for continuous random variable, there exists a function f(x) similar to $F(X) = P_x\ (x \leq x) = \int^x_{\infty} f\ (y)dy$ – this is known as probability density function. It satisfies the following:

i. $f(x) \geq 0$ for all x
ii. $\int_a^{\infty} f\ (x)\ dx = 1$
iii. $P\ (X < a) = \int_a^{\infty} f(x)\ dx = 1 - \int_a^{x\infty} f\ (x)dy = 1 - P\ (X > a)$
iv. A is any interval, $P\ (X\ (A) = \int_a (f(x)\ dx$

The symbol ∫ means sign of integration. Its principle involves:

If $y = x^{n-}$
$\int^b_a ydx = \int x^n dx$

$$= \frac{x^{n+1}}{n+1}\Big)^b_a$$

$$= \frac{b^{n+1}}{n+1} - \frac{a^{n+1}}{n+1}$$

$$= \frac{1}{n+1}(b^{n+1} - a^{n+1})$$

Example 3

The Probability Density Function (PDF) of a random variable *x* is given as:

$$F(x) = \begin{cases} 2x & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Shows that f(x) is a true *pdf.*

Solution

Checking the validity of (i) and (ii)

i. $f(x) \geq 0$ in the interval (0, 1)

ii. $\int^1_0 f(x)\,dx = \int^1_0 2x\,dx$

$= 2\int^1_0 x\,dx$

$= 2\left\{\frac{x^{1+1}}{1+1}\right\}^1_0$

$= 2\left\{\frac{x^2}{2}\right\}^1_0$

$= 2(\tfrac{1}{2}) + 2(\tfrac{0}{2})$

$= \tfrac{2}{2} + 0$

$= 1$

Thus, *f(x)* is true *pdf.*

3.4 Expectation and Variance of a Continuous Random Variable

If *x* is a continuous random variable with pdf (f(x) for $-\infty < x < \infty$, $\sum(X)$ in this case, is given as:

$\sum(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$

The variance of a continuous random variable with pdf f(x), for $-\infty < x < \infty$, is given $V(x) = \int_{-\infty}^{\infty} X^2 f(x)\, dx - \left(\int_{-\infty}^{\infty} dx \right)^2$

Example 4

Let *x* be random variable with density function.

$$F_x(x) = \begin{cases} 2x \text{ for } 0 \; x < 1 \\ 0 \quad \text{otherwise} \end{cases}$$

Find (a) E(x) (b) var (x)

Solution

$E(x) = \int_0^1 x\, f(x)\, dx$
$= \int_0^1 x\, 2x dx$
$= \int_0^1 x\, 2x^2 dx$
$= 2\int_0^1 x\, 2^2 dx$
$= 2 \left\{ \frac{x^{2+1}}{2+1} \right\}_0^1$
$= 2 \left\{ \frac{x^3}{3} \right\}_0^1$

$= 2 ({}^1/_3) + 2 ({}^0/_3)$
$= {}^2/_3$

$V(X) = E(x^2) - \{E(X)\}^2$
$\int_0^1 x\, f(x)\, dx - (\int_0^1 x\, f(x)\, dx)^2$
$E(x^2) = \int_0^1 x^2 (2x)\, dx$
$= \int_0^1 2 x^2 (2x)\, dx = 2 \int_0^1 x^3\, dx$
$= 2\, x^4 \; {}^1 = 2\,(1/4) - 2\,({}^0/_4)$
$= ½$

$V(x) = ½ - ({}^2/_3)^2$
$= ½ - {}^4/_9 = \frac{9-8}{18} = \frac{1}{18}$.

3.5 Some Standard Discrete Distribution

Let us consider the following

a. The Bernouli

Any experiment with two possible outcomes is called Bernoulli experiment. The distribution function is written as-

$P(x) = P^x (1-P)^{1-x}$, $X = 0, 1$

The expectation of Bernoulli distribution is very easy to get.

$$\begin{aligned} E(X) &= \sum X \Pr(X) \\ &= 0 (P^0) (1-P)^{1-0} + 1 P^1 (1-P)^{1-1} \\ &= 0 + P (1-P)^0 \\ &= P \end{aligned}$$

Similarly, it is easy to compute the variance from the axiom that:

$$V(X) = EX^2 - (E(X))^2$$

$$\begin{aligned} E(x^2) &= \sum X^2 \Pr(X) \\ &= (0) P^0 (1-P)^{1-0} + 1^2 P^1 (1-P)^{1-1} \\ &= 0 + p (1-P) \\ &= P^0 \end{aligned}$$

$$\begin{aligned} V(X) &= P - (P)^2 \\ &= P (P-P)^2 \\ &= P (1-P) \\ &= Pq \end{aligned}$$

Example 5

Given X ~ Ber (0.05), find E(x) and V(x)

Solution

$E(X) = P = 0.5$

$V(X) = Pq = 0.5 = 0.5 \times 0.5 = 0.25$

b. The binomial random variable

A common source of business data is an opinion or preference survey. Many of these surveys result in dichotomous response i.e., responses that admit one or two possible alternatives, such as *Yes* or *No*. the number of *Yes* response (or *No* response) will, usually, have a binomial probability distribution. For example, suppose a random sample of customers is selected from the totality of potential consumers of a particular product; the number of consumers in the sample who prefer

the product to its competition is a random variable that has a binomial probability distribution.

Characteristics of a binomial random variable

1. The experiment consists of identical trials.
2. There are only two possible outcomes on each trial; we denote one outcome by *S* (for success) and the other by *f* (for failure).
3. The probability of *S* remains the same from trial to trial. This probability is denoted by *P*, and the probability of *F* is denoted by q. Note that P + q = 1.
4. The trials are independent.
5. The binomial random variable *X* is the number of *S's* in *n* trials.

The probability distribution is given as-

$P(X) = \binom{n}{x} P^x q^{n-x} (x = 0, 1, 2, \ldots, n)$

Where P = Probability of a success on single trial.

q = 1 – P
n = number of trials

X = number of success in trials
$\binom{n}{x} = \frac{nl}{(n-x)!\,x!}$.

b. The mean and variance of a binomial distribution

E (X) = μ = np
V(X) = npq
√(x) = √npq

Example 6

A sales person has found, over a long period of time, that the probability of making a sale by calling on a customer is 0.5. If this sales person calls on 5 customers in a day, find the probability of making (a) exactly 3 sales (b) 3 or more sales (c) fewer than 3 sales (d) no sales (e) 5 sales.

Solution

$P(X) = \binom{n}{x} P^x q^{n-x} (x = 0, 1, \ldots, n)$
$= 0.5 \quad q = 1 - 0.5 = 0.5$
(a) $P(x = 3) = \binom{5}{3} (0.5)^3 (0.5)^{5-3}$
$= \binom{5}{3} (0.5)^3 (0.5)^2$

= 10 x 0.125 x 0.125
= 0.15635

(b) $P(X \geq 3)$ = $1 - P(x \leq 2)$ or
= $P(x = 3) + P(x = 4) + P(X = 5)$

Let us use the first

= $1 - P(\leq 2)$ or
= $1 - \binom{5}{0}(0.5)^0 (0.5)^5 + \binom{5}{p}(0.5)^1 (0.5)^4 + \binom{5}{2}(0.5)^2 (0.5)^3$
= $1 - (0.03125 + 5(0.05)^3 + 10(0.125)(0.125)$
= $1 - (0.03125) + 0.15625 + 0.3125)$
= $1 - 0.5$
= 0.5

(c) $P(X < 3)$ = $P(x = 0) + (X = 1) + P(x = 2)$
= 0.03125 + 0.15625 + 0.3125

(d) $P(X = 0)$ = 0.03125

(e) $P(X = 5)$ = $\binom{5}{5}(0.5)^5 (0.5)^{5-5}$
= $1 \times (0.5)^5 (0.5)^0$
= 0.03125

Note that when *n* becomes large, binomial Probability Mass Function (PMF) may exert pressure on the computation as it becomes cumbersome. Fortunately, probabilities for different values of *n, p, and x* have been tabulated, so that instead of calculating the probabilities, we may consult a table to obtain the desire result.

3.6 Chebyshev's Theorem

Given the probability distribution of a random variable *x*, with mean μ and standard deviation σ, the probability of observing a value of x within K standard deviation of μ is at least $1 - \frac{1}{K^2}$.

Example 7

It is stated that 65% of a certain population prefers a particular brand of toothpaste, suppose we interview a random sample of 500 people from this population; within what interval would we expect the number of success out of these 500 trials to lie with a probability of 0.95.

Solution

From Chebyshev's theorem

$1 - \frac{1}{K^2} = 0.95$

$\frac{K^2 - 1}{} = 0.95$

K^2

$K^2 - 1 = 0.95k^2$

$K^2 = 0.95k^2 = 1$

$K^2 (1 – 095) = 1$

$K^2 (0.05) = 1$

$K^2 = \frac{1}{0.05}$.

$K^2 = 20$

$K = \sqrt{20} = 4$

We will say that the probability is, at least, 0.95- that the number of success we would observe is with 4 standard deviation of the mean-

So $\mu = np = 500 \times 65\%$

$= 500 \times 0.65$

$= 325$

$r = \sqrt{npq}$

$= \sqrt{np(1-p)} = \sqrt{500(0.65)(0.35)}$

$= 10.67$

Since 4 (10.67) = 42.68, we find the interval we seek to be

$\mu \pm \sigma = 325 \pm 42.68$

= (282.32, 367.68)

Suppose that we find only 250 out of the 500 who prefer the brand of toothpaste; we may question the truthfulness of the statement that 65% of the population prefer the brand, since 250 is more than 5 standard deviations from the mean.

Chebyshev's theorem tells us that the probability of this occurring,is equal to-

$\frac{1}{K^2} = \frac{1}{4^2} = \frac{1}{14} =$ 0.06 or less

3.7 The Poisson Distribution

A type of probability distribution useful in describing the number of events that will occur in a specific period of time in a specific area or outcome is the Poisson distribution (named after the eighteen – century physicist and mathematician Simeon Poisson). The following are typically examples of random variable for which Poisson distribution provides a good model.

1. The number of industrial accidents in a given manufacturing plant, per month, as observed by a plant safety supervisor.

2. The number of noticeable surface defects (scratches etc.) found by quality inspectors on a new automobile (or any product from a manufacturing plant).
3. The parts- per million, of some toxicants found in water or air emission from a manufacturing plant (a random variable of great interest to both business community and environmental protection agency.
4. The number of arithmetic errors per 100 invoice (or per 1,000 invoices) in the accounting records of a company.
5. The number of customer arrivals per unit time at service desks (a service station, hospital, bank, a supermarket etc.
6. The number of death claims- per day, received by insurance company.
7. The number of breakdown of an electronic computer per month etc.

3.7.1 Characteristics of a Poisson Random Variable

1. The experiment consists of counting the number of times a particular event occurs during a given unit of time or in a given area or volume (or weight, distance or any other unit of measurement.
2. The probability that an event occurs in a given unit of time (area or volume is the same for all the units).
3. The number of events that occurs in one unit of time, area or volume is independent of the number that occurs in the other units.
4. The mean (or expected) number of events in each unit will be denoted by the Greek letter LAMBDA, λ
5. Poisson's probability distribution also provides a good approximation to a binomial distribution with mean $\overline{X}$ = np

The probability distribution is given as:

$$P(x) = \frac{\lambda^{x} - \lambda^{-x}}{X!} \quad (x = 0, 1, \ldots)$$

The mean or expected value $\mu = \lambda = \sigma^2 = \lambda$

Where λ = mean number of events during a given time period, or over a specific area or volume.

e = 2 – 71828

3.7.2 Poisson's Distribution as an Approximation to the Binomial Distribution

We can use Poisson's distribution to approximate the binomial distribution when *n* is large- $n \geq 20$ and P is small- $P \leq 0.05$. Where these conditions are met, we can express the relationship as follows.

$({}^{n}_{p})\, q^{n-x} P^{x} \qquad \frac{e^{-np}(np)^{x}}{x!}$ where $\lambda - np$

Example 8

The number of defects square foot of a certain manufactured fabric is distributed using Poisson's model- x = 0.08. If a square foot of fabric is inspected, what is the probability that the number of defects observed is (a) zero (b) at least 1 (c) exactly 2?

Solution

(a) $p(x = x) = \frac{x^{x} e^{->}}{x!}$ $(x = 0,1 \ldots)$

$X = 0.08$

$P(x = 0) = \frac{(0.08)^{0} \times 2.718^{-0.08}}{0!}$

$= \frac{1}{2.718^{0.08}} = \frac{1}{1.083} = 0.923$

(b) $P(x \geq 1) = 1 - Pr(x \leq 0)$

$= 1 - 0.923$

$= 0.077$

(c) $P(X = 2) = \frac{(0.08)^{2} e^{-0.08}}{21}$

$= \frac{0.08}{21 (e^{0.08})}$.

$= \frac{0.08}{2 \times 1.083}$.

$= \frac{0.08}{2.166}$.

$= 0.0369$

$= 0.037$

3.8 Hyper Geometric Random Variable

The hyper geometric probability distribution provides a realistic model for some types of enumerative (count) data. The characteristic of hyper geometric distribution are as listed below.

1. The experiment consist of randomly drawing *n* elements without replacement from a set of *N* element, *r* of which S's (for successful and (N – r) or which are F's (for failure).
2. The hyper geometric random variable x is the number of S's in the draw of n elements.

3.8.1 Probability Distribution, Mean and Variance of Hyper Geometric Random Variable

The probability distribution of the hyper geometric random variable is as follows:

$$P(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

(x = maximum (0, n – (N – r), … minimum (r, n)
The mean and variance are, respectively-

$$\mu = \frac{nr}{N} \quad , \quad \sigma^2 = \frac{r(N-r)n(N-n)}{N^2(N-1)}$$

Example 9

Given that x is a hyper geoetric random variable with N = 8, n = 3, and r = 5, and r = 5, compute (a) P (x = 1) (b) P (x = 0) (c) P (x $\geq$ 4) (d) compute the μ and σ

$$P(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

r = 5, N = 8, n = 3

(a)

$$P(x=1) \quad \frac{\binom{5}{1}\binom{8-5}{3-1}}{\binom{8}{3}} = \frac{\binom{5}{1}\binom{3}{2}}{\binom{8}{3}}$$

$$= \frac{5 \times 3}{56} = \frac{15}{56}$$

(b)

$$P(x=0) \quad \frac{\binom{5}{0}\binom{8-5}{3-0}}{\binom{8}{3}} = \frac{1 \times 1}{56} = \frac{1}{56}$$

(c) $P(x \geq 4) = 1 - P(x \leq 3)$

$= 1 - (P(x=0) + P(x=1) + P(x-2) + P(x=3)$

$$P(x=2) = \frac{\binom{5}{2}\binom{8-5}{3-2}}{\binom{8}{3}} = \frac{1 \times 1}{56} = \frac{1}{56}$$

$$= \frac{\binom{5}{2}\binom{3}{0}}{\binom{8}{3}} = \frac{10 \times 3}{56} = \frac{15}{56}$$

$$P(x=3) = \frac{\binom{5}{3}\binom{3}{0}}{\binom{8}{3}} = \frac{10 \times 1}{56} = \frac{5}{56}$$

$$P(x \geq 4) = 1 - \left(\frac{1}{56} + \frac{15}{56} + \frac{15}{28} + \frac{5}{28}\right)$$

$$= 1 - \left(\frac{1}{56} + \frac{15}{56} + \frac{30}{56} + \frac{10}{56}\right)$$

$= 1 - 1$

$= 0$

$$\mu = \frac{3 \times 5}{8} = \frac{15}{8}$$

$$\sigma^2 = \frac{5(8-5)\,3(8-3)}{8^2(8-1)}$$

$$= \frac{5(3)(3)(5)}{64 \times 7}$$

$$= \frac{245}{448}$$

(d) $= 0.502$

3.9 Geometric Random Variable

Another discrete random variable that has many business applications is the geometric random variable. Like the binomial random variable, it arises naturally from a discussion of a coin-tossing experiment (whether the coin is balanced or unbalanced). But instead of tossing the coin a fix number of times and observing the number x, of heads, we toss the coin and count the number, x, tosses until the first head appears.

The following are the characteristics of the geometric random variable.

(1) The experiment consists of a sequence of independent trials.
(2) Each trial results in one of two outcomes. We denote one of them by *s* and other by *F*.
(3) The probability of *S* remain the same from trial to trial. We denote this by P.
(4) The geometric random variable x is defined to be the number of trials until S is selected.

The probability distribution of geometric random is given as:

$P(x) = q^{x-1}P$, $(x = 1, 2, 3, 4, \ldots)$
The mean and variance are:

$$\mu = \frac{1}{p}, \quad \sigma^2 = \frac{q}{P^2}$$

Example 10

Given that x is a geometric random variable with P = 0.2, compute the following: (a) p (x = 1) (b) p (x = 2) (c) $x \leq 3$) (d) find the μ and σ

Solution

$P(x) = q^{x-1}$ $(x = 1, 2, 3, 4, \ldots)$
$P = 0.2$ $q = (1 - 0.2) - 0.8$

(a) $P(X - 1) = (0.8)^{2-1}(0.2)$
$= 0.8 \times 0.2 = 0.16$

(b) $P(x \leq 3) = P(x = 1) + P(x = 2) + P(X = 3)$
$P(x = 3) = (0.8)^{3-1}(0.2)$
$= (0.8)^2 \times (0.2)$
$= 0.128$

(c) $P(x \leq 3) = 0.128$
$= 0.2 + 0.16 + 0.128$
$= 0.488$

3.10 Some Standard Continuous Probability Distribution

Here, you will be exposed to the following continuous probability distribution.

(a) The uniform distribution
(b) Exponential distribution
(c) Normal distribution

a. The uniform distribution

Perhaps the simplest of all the continuous probability distributions is the uniform distribution. The frequency function has a rectangular shape- as shown in the figure below. Note that the possible value of *x* consists of all points on the real line between point *c* and *d*; the height of f(x) is constant in the interval and equals $^{1}/(d-c)$.

Therefore, the total area under *f(x)* is given by -

Total area of rectangle = BXH = $(d - c)\left(\frac{1}{d - c}\right) = 1$

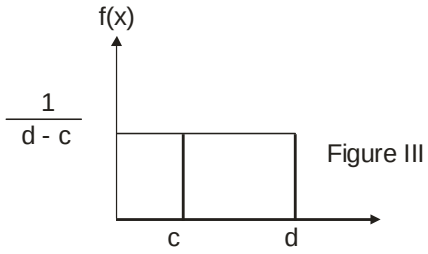


Figure III

Figure 4.3

The efficiency of the uniform distribution is that it provides a model for continuous random variable that is evenly distributed over a certain interval. That is, there is no clustering of values around any value; instead, there is an even spread over the entire region of possible value.

The probability distribution is given as:

$F(x) = \frac{1}{d - c}$ $(c \leq x < d)$ while the mean and variance are:

$$\mu = \int_c^d x\ f(x)dx = \int_c^d x\ \frac{1}{(d - c)}\ dx$$

$$\mu = \frac{c + d}{2}$$

$$V(x)\ \ E(x^2) - (E(x))^2$$

$$= \frac{1}{d-c}\int_c^d x^2\,dx - \left(\frac{1}{d-c}\int_c^d x\,dx\right)^2$$

$$= \frac{(d-c)^2}{12}$$

Example 11

A research department of a steal manufacturing company observes that one of the company's rolling machines is producing sheets of steel of varying thickness. The thickness is a uniform random variable, with values between 150 and 200 millimetres. Any sheets less than 160 millimetres thick must be scrapped because they are unacceptable to buyers.

(a) Calculate the mean and variance of the thickness of the sheet produced by this machine.
(b) Calculate the fraction of steel sheets produced by this machine that have to be scrapped.

Solution

(a) $\mu = \frac{c+d}{2}$, $c \leq x \leq d$

let x be that thickness.

C = 150 d = 200

$$\mu = \frac{150 + 200}{2} = 175 \text{ millimeters}$$

$$\sigma = \frac{(d-c)}{2} = \frac{(200-150)}{12} = 28.3\text{mm}$$

(b) P(x < 160) = (Base) (Height) = (160 – 150) f(x)

$$= 10\left(\frac{1}{200-150}\right)$$

$$= \frac{10}{50} = {}^1/_5$$

That is, 20% of all the sheets made by the machine must be scrapped (See diagram below).

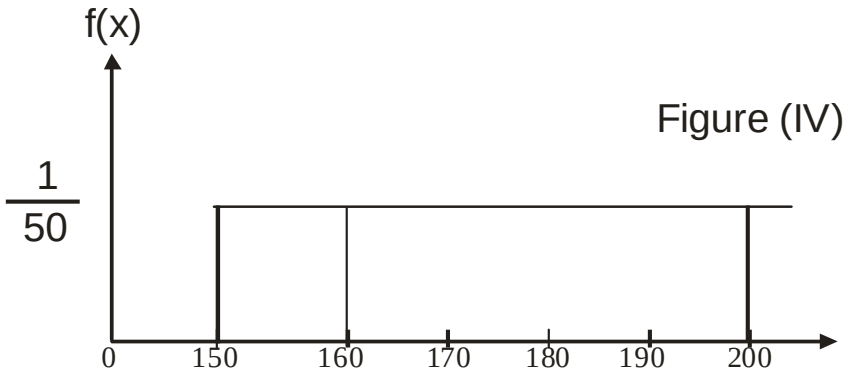


Figure (IV)

Fig. 4.4

b. Exponential distribution

Another important probability distribution that is more beneficiary in describing business data is exponential distribution. Two business phenomena with frequency functions that may be well approximated by exponential distribution are the length of time between arrivals at a fast food centre, and the length of time between the filing of claims in a small insurance office. Note that in each of these examples, the measurements are the length of time between certain events. For this reason, the experimental distribution is sometimes called the waiting time distribution.

The probability distribution of exponential distribution is given as-

$f(x) = \lambda e^{-\lambda x} \quad (X > 0)$

The graph below shows experimental distribution when $\lambda = 0.5,\ x = 1$ and 2

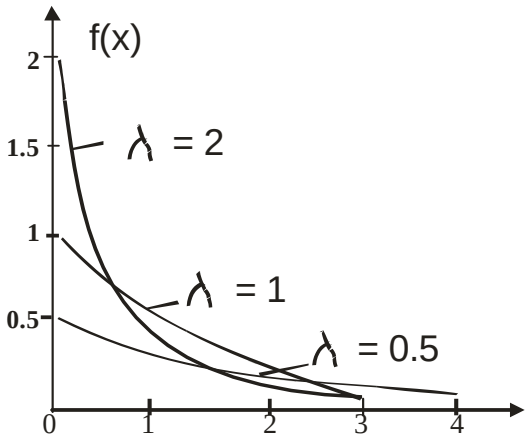


Fig. 4.5

The mean and variance are $\mu = \lambda$ and $\sqrt{}^2 = \lambda^2 = \sqrt{}^2 = \lambda$

To find the area, *A* to the right of *A* number, "a" for an exponential distribution involve the use of integration by substitution. That is-

$A = P(x \geq a) = e^{-\lambda a}$

$$\int_{-\infty}^{a} f(x)dx = \int_{a}^{\infty} \lambda e^{-\lambda a} dx$$

Let $u = \lambda x$

$du = \theta \lambda dx$

$dx = \frac{du}{-\lambda}$

$$\int_{a}^{\infty} \lambda e^{-\lambda a} dx = \int_{a}^{\infty} \lambda e^{u} \frac{du}{-\lambda} = -\int_{a}^{\infty} e^{u} du = -\left(e^{u}\right)_{a}^{\infty}$$

$$= 1\left(e^{-\lambda a}\right)_{a}^{\infty} = -e^{-\lambda(\infty)} - (-)e^{-\lambda(a)}$$

$= e^{-\lambda a}$

Therefore, to find *P* ($x \geq a$) to left of "a" we use the transformation.

$1 - P(x \geq a)$

Example 12

The length of time (in days) for the sales of an automobile is modelled as an exponential random variable with x = 0.5. What is the probability that the salesperson gives:

(a) more than 5 days without a sale?
(b) less than 5 days without a sale?

Solution

a. Since a = 5
$A = e^{-\lambda a}$
$A = e^{(0.5)5} = e^{-2.5}$

A = 0.082085

That is, our model indicates that the automobile sales person has a probability of about 0.08 of going more than 5 days without sale.

b. $P(x \leq 5) = 1 - P(x \geq 5)$
$= 1 - 0.08$
$= 0.92$

That is, our model indicates that 0.92 of the salesperson going less than 5 days without a sale.

c. The normal distribution

We now come to the most important distribution in statistics and probability. This is because many practical measurement or problems can be explained by using the normal distribution. More over, most of the other probability distributions can be approximated by the normal distribution.
The normal distribution function is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \quad e - \frac{(x-u)^2}{\sigma^2}$$

Where:
μ = mean of normal random variable, x
π = 3.142
e = 2.71828.

The formula, as you can see above, is complex. The expression is stated here, but unfortunately, it is never necessary to use it in practical because it is possible to solve all relevant problems using a single table of probability for the so-called standard normal distribution. So whenever we take the integral of this function it results into the area under the normal curve.

$$A = \int_a^{\infty} f(x)dr \qquad = \int_a^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \quad e\frac{1}{2}\frac{(x-u)^2}{\sigma^2}$$

$$P(z) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \quad e\frac{1}{2}(z)^2$$

$$\text{Where } (z) = \frac{x \quad u}{\sigma}$$

Equation is called the standard normal distribution. The normal distribution shape is shown below.

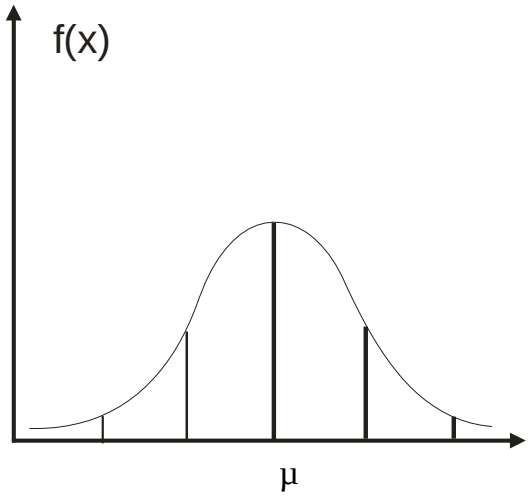


Fig. 4.6

Properties of normal distribution

1. It is symmetrical about its mean.
2. The mean, mode and median are all equal.
3. The total area under the curve above the x – axis is equal to 1.
4. It is a bell shape.
5. 50% of the area is to the right of the mean, and 50% is to the left.

68%, 95% and 99.7% of the area under the normal curve lies within 1σ, 2 σ, 3 σ or the mean respectively (see the figure below).

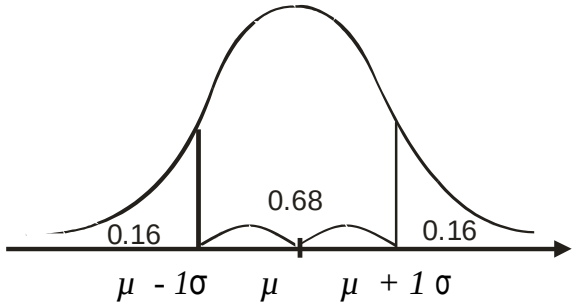


Fig. 4.7

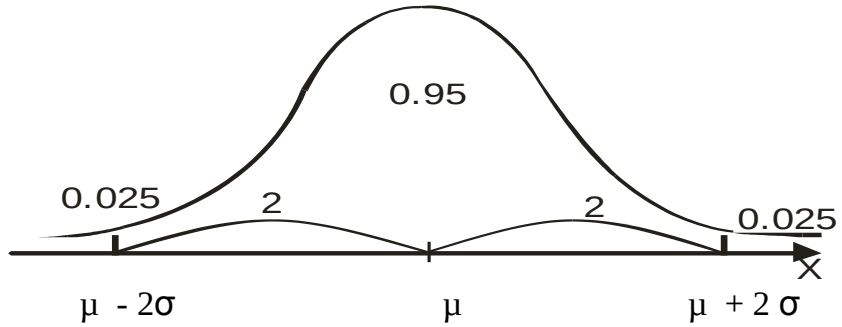


Fig. 4.8

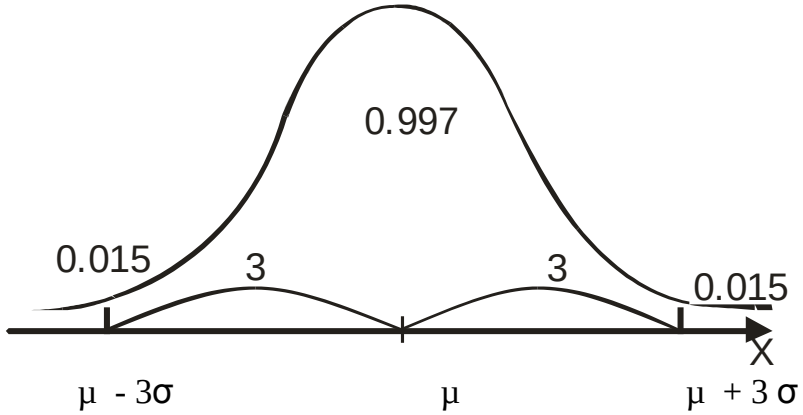


Fig. 4.9

Properties of standard normal distribution

(1) It has mean 0
(2) Its variance 1

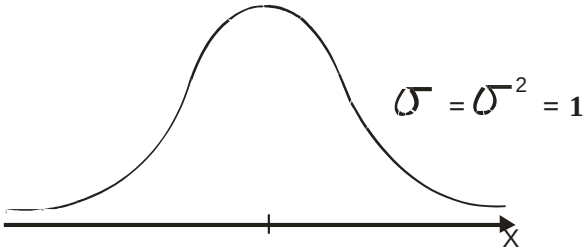


Fig. 4.10

Example 13

(1) Given the standard normal distribution, find:

(a) the area under the curve between Z = 0 and Z = 1.54
(b) P ($\underline{Z} \geq 0.65$)
(c) P ($Z < 2.33$)
(d) P ($-1.96 \; \underline{Z \geq} \; \underline{\leq} 1.96$)

(2) Given a normal distribution population of values with a mean of 76 and a standard deviation of 10:

(a) what proportion of value is between 71 and 82?
(b) what proportion is greater than 75?
(c) what is the probability that a value, picked at random from this population, is less than 78?

Solution

Standard normal distribution

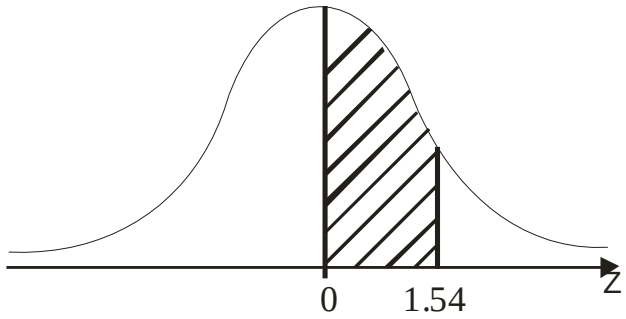


Fig. 4.11

Locating z = 1.54 in the normal distribution table gives 1.5 under 0.04382 = 0.4382
= 43.82%

(b)

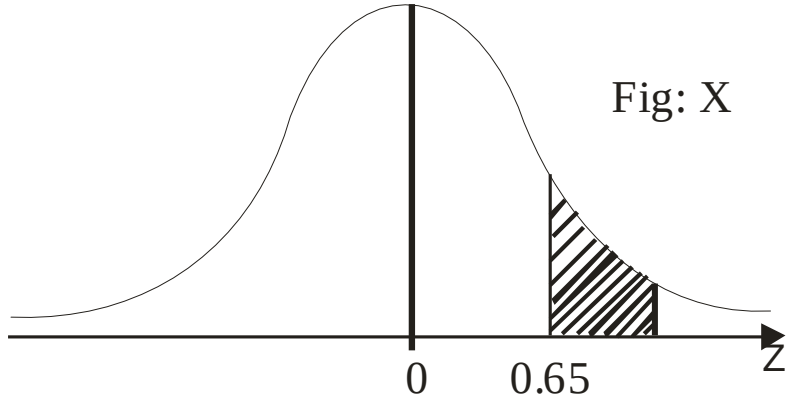


Fig. 4.12

Location 0.6 under 5 from the normal distribution table gives 0.2422. we obtain the area by subtraction of the area between 0 and 0.65 from 0.5, i.e $P(2 \leq 0.5) = 0.5 - 0.2422 = 0.2578$.

(c)

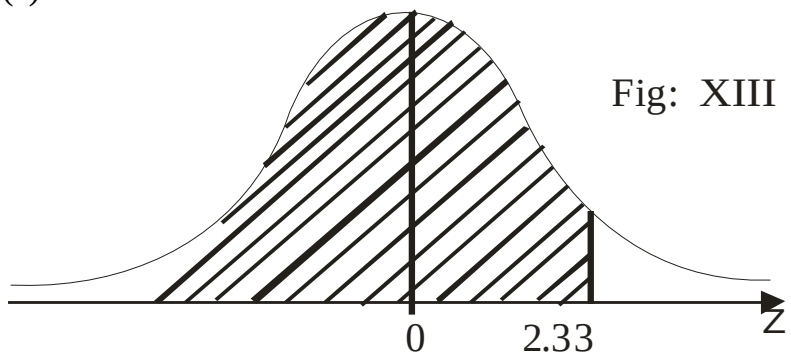


Fig. 4.13

Locate 2.3 under 3, it gives 0.4901 to get P($\geq$ 2.33) we add 0.5 to 0.49901
i.e 0.5 + 0.4901
= 0.9901
(d)

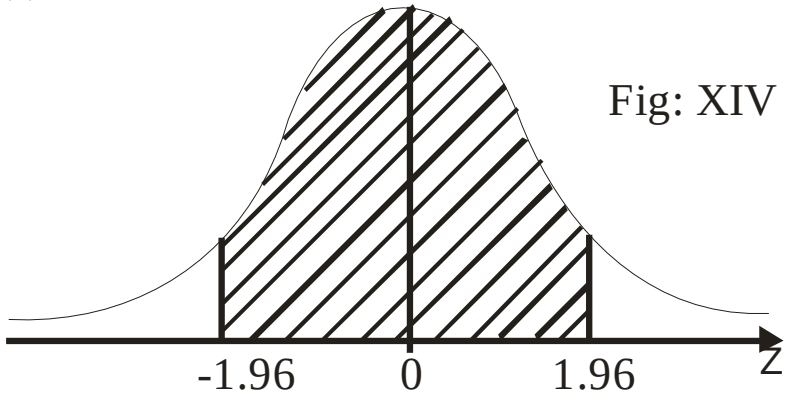


Fig: XIV

Fig. 4.14

Locate 1.9 under 6 we get 0.4750 because of symmetry, we know that the area between 0 and -1.96 is the same as the area between 0 and 1.96. Therefore, we double the desired result.

P(-1.96 $\leq$ Z $\leq$ 1.96) = 2 x 0.4750
= 0.95
(2)

(a) $Z = \frac{X - U}{\sigma}$

$X_1 = 71$ and $X_2 = 82$, $\mu = 76$, $\sigma = 10$

$P(\frac{X_1 - U}{\sigma} < > < \frac{X_2 - U}{\sigma}) = P(\frac{71 - 7}{10} \leq Z \frac{82 - 76}{10})$

= P(-0.5 $\leq$ Z 0.6)

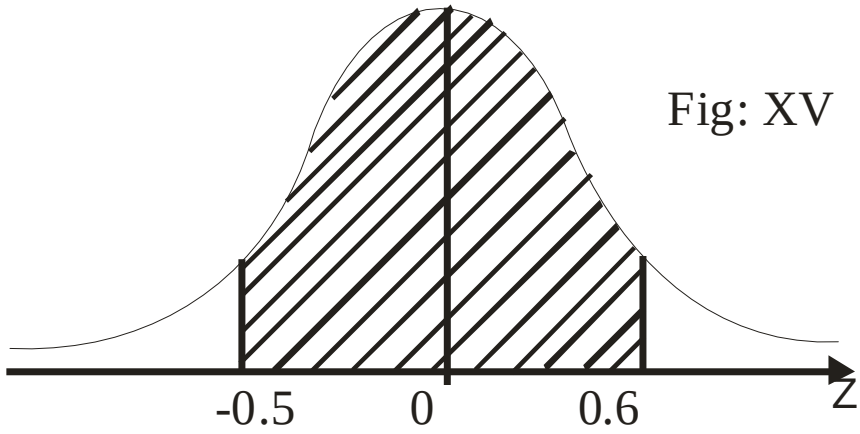


Fig: XV

Fig. 4.16

$P(-0.5 \leq Z \leq 0.06)$
$= 0.01915 + 0.2257$
$= 41.72\%$

(b) $P(Z > \frac{X - U}{\sigma}) = P(Z > \frac{75 - 76}{10})$

$= P(Z > 0.1)$

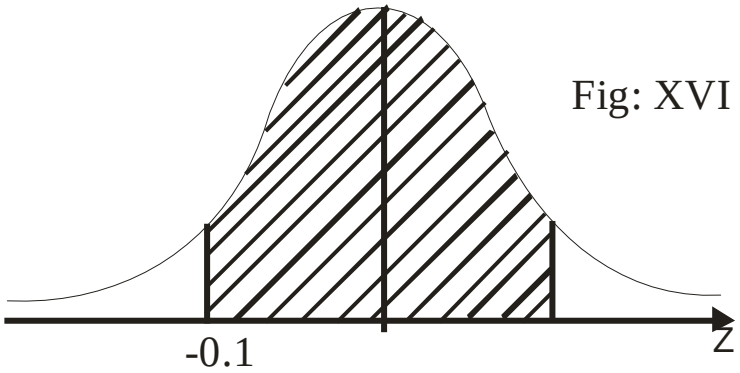


Fig. 4.17

$P(Z > 0.1) = 0.5 + 0.0398$
$= 0.5398$
$= 53.98\%$

(c) $Pr(Z < \frac{78 - 76}{10}) = P(Z < 0.2)$

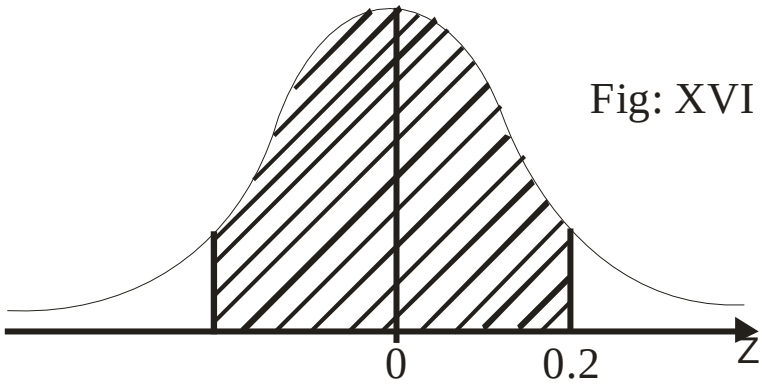


Fig. 4.18

$Pr(Z < 0.2) = 0.5 + 0.0793$
$= 0.5793$

3.10.1 Normal Approximation to the Binomial Distribution

A rule of thumb that is frequently followed states that normal approximation to the binomial is appropriate when np and n (1 – P) are both greater than 5. Then, the transformation becomes-

$$Z = \frac{X - nP}{\sqrt{nP(1 - P)}}$$

The result of the transformation above is appropriate with little modification as we allow x to be adjusted by $\pm$ 0.5 within the normal interval, N (0,1). This adjustment is often called the continuity correction. This implies that:

(i) $P(X = a) = P(a - \frac{1}{2} \leq X \leq a + \frac{1}{2})$
(ii) $P(a \leq X \leq b) = P(a - \frac{1}{2} \leq X \leq b + 1/2)$

Example 14

Given that n = 20 and P = 0.3 find the P ($5 \leq X \leq 10$)

Solution

nP = 20 x 0.3 = 6 and n(1-P) = 20 x 0.7 = 14.
Since they are greater than 5, we make use of the normal approximation to the binomial distribution.

$$P(5 < x < 10) = P\left(\frac{5 - 0.5 - 6}{\sqrt{6(0.7)}} \leq Z \leq \frac{10 + 0.5 - 6}{\sqrt{6(0.7)}}\right)$$
$$= P(-0.73 \leq Z \leq 2.20)$$

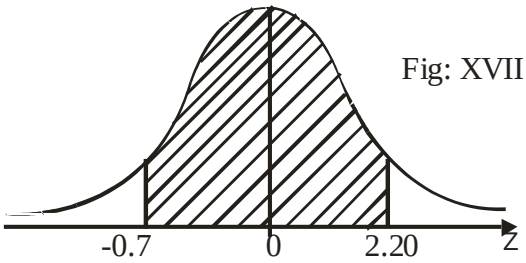


Fig: XVIII

Fig. 4.19

$$P(-0.73 \leq Z \leq 2.20) = 0.2673 + 0.4861$$
$$= 0.75534$$

SELF-ASSESSMENT EXERCISE

In a certain large firm, 30% of the employees are female. A random sample of 50 is selected from this population. What is the probability that the number of female will be between 20 and 24, inclusive.

4.0 CONCLUSION

In this unit, you would have noticed that probability distribution is an essential aspect of statistics because it will enable you to overcome the fear associated with measurements. You can only measure accurately when you are familiar with various variables with their respective probability distributions.

5.0 SUMMARY

By now, your knowledge of probability distributions would have being enhanced. You can now construct mathematical model which accurately describes the situation associated with a particular event in which you are interested. That is, it is either you are dealing with the counting of events, which are discrete or you are dealing with continuous variables which are associated with measurements such as heights, weights, etc.

6.0 TUTOR-MARKED ASSIGNMENT

i. If two fair dice are tossed:

   (a) describe the sample space
   (b) find the probability that the sum of the numbers on the upturned faces of the dice is less than 9.
   (c) find the probability that number on the second die is less than 3 if the number on the first die is greater than 3?
   (d) if the number on the first die is 2, what is the probability that the number on the second is an odd number.

ii. Given a normal distribution of values with mean of 120 and a variance of 16, what proportion of the values are greater than 14?

7.0 REFERENCES/FURTHER READING

Brian, P.M. & Philip, M.N. (2006). *Applied Statistics for Public Policy*. New Delhi: Prentice-Hall.

James, T. & Benon, P.G. (1988). *Statistics for Business and Economics*.(4th ed.). United States: Dellen Publishing Company Inc.

Kasumu, R. B. (2002). *Introduction to Probability Theory. A First Course*. JAB Publishers.

MODULE 3

UNIT 1 TEST OF HYPOTHESIS

CONTENTS

1.0 INTRODUCTION

Now that you have learnt about probability distribution in the earlier unit, a general fame work for testing hypothesis will be provided- as a core aspect of statistical inference. In broad outline, you will be exposed to various possible hypotheses that are common to behavioural sciences in the light of relevant data which have been collected, in order to arrive at probable administrative decision. This topic is of central importance in statistics because hypothesis tests are widely used as the basis for making decisions in the industrial sector, government and in research more generally.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain the principles of statistical inference
- describe the basic concepts of statistical inference and the terminologies used in conducting tests of hypothesis
- identify various types of hypotheses testing.

3.0 MAIN CONTENT

3.1 Statistical Hypothesis

A statistical hypothesis is an assertion or statement about a probability distribution, or a population parameter**.**

3.2 Null Hypothesis

Null hypothesis (H_0) has to do with the theory about the values of one or more population parameter. The theory, generally, represents the status quo, which we accept until proven false.

3.3 Critical Region or Rejection Region

The numerical values of the statistics of the test on the basis of which the null hypothesis will be rejected. The rejection or critical region is chosen so that the probability is α that it will contain the test statistic when the null hypothesis is true.

3.4 Alternative Research Hypothesis (H_a)

A theory that contradicts the null hypothesis; the theory, generally, represents that which we will accept only when sufficient evidence exists to establish the truth.

3.5 Acceptance Region

This is the region in the sample space which leads to acceptance or H_0.

3.6 Type 1 Error

This is the act of rejecting a true hypothesis.

3.7 Level of Significance

It is denoted as α, it is the level at which the researcher will be able to risk type 1 error. It is often specified e.g. $\Phi = 0; 0.05, 0.01$ etc.

3.8 Type II Error

This is the act of accepting a false hypothesis. The type II error is denoted as β.

3.9 Two-Tailed and One-Tailed Tests

An alternative to a null hypothesis determines whether a test of hypothesis will be at both sides of the normal course, which we called two-tailed test. It is stated as:

$H_a\ ^0\mu \neq {}^\mu{}_0$ meaning $\mu < \mu_0$ or $\mu > \mu_0$.

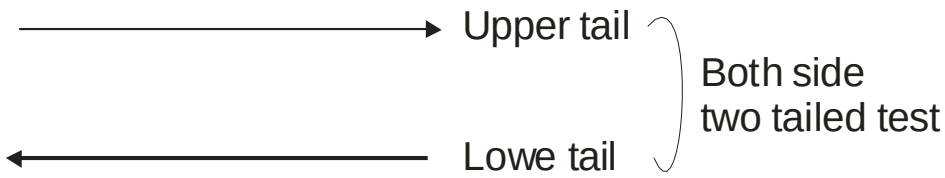


Fig. 1.1

While H_a: $U > U_0$, meaning $U > U_0$ – One-tailed upperside
And H_a: $U < U_0$, meaning $U < U_0$ – One lower tailed

3.10 Procedure for Hypothesis Testing

Let us consider the following.

1. Statement of the hypothesis
2. Identification of the test statistic and its distributions
3. Specification of the signification level
4. Statement of the decision rule
5. Collecting the data and doing the calculations
6. Making the statistical decision
7. Making the administrative decision

We may compress these seven steps above into five steps:

1. state the hypothesis
2. specify the significance level
3. test the statistics
4. state the critical value
5. make conclusion.

3.11 One Population Mean

The mean of a normally distributed population - Known as population variance.

Example 1

The mean number of accidents experienced by a group of 200 coal miners, over a period, was 2.0- with standard deviation 1.487. Test the hypothesis that the mean number of accidents, amongst all miners represented by this sample is 1.75; let $\alpha = 0.05$

To construct the test, we follow the 5 steps faithfully.

Step I: H_0 $\mu = 1.75$ $H_i = \mu \neq 1.75$ (two-tailed test)
Step II: Let $\alpha/_2 = 0.05/_2$
Step III: Z $\alpha = 0.05 = 0.025 = \pm 1.96$

Step IV: $Z = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$.

$\overline{X} = 2$, $\sigma = 1.487$, $n = 200$
$M = 175$

$$Z = \frac{2 - 1.75}{\dfrac{1.487}{\sqrt{200}}} = \frac{0.25}{1.487} \times \sqrt{200}$$

$= 2.38$

Step V: Since $Z_{cal} > Z_{critical}$ - rejected the H_0 ,and concluded that mean is not 1.75.

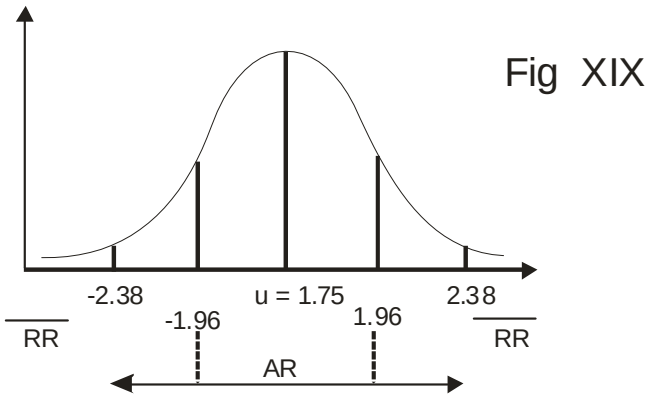


Fig. 1.2

The *P* value can be calculated thus:

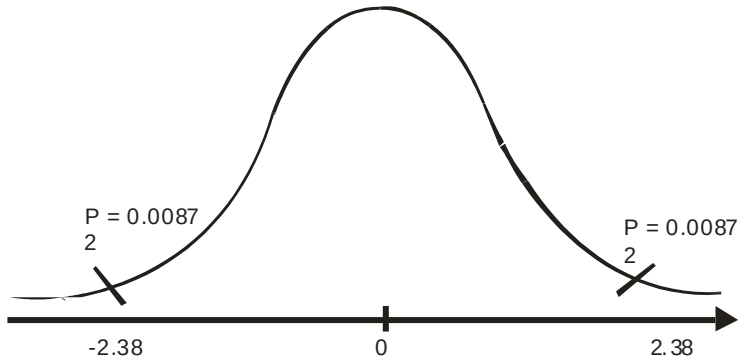


Fig. 1.3

From the above 2.3 under 8 is 0.4913; area left of the curve is 0.5 – 0.4913, and right of the curve is 0.5 - .4913.

P = 2 (0.0087)
P = 0.0174

Since 0.0174 is less than the chosen significance level, we reject H_0

3.12 One Population Mean/The Mean of a Normally Distributed Population (Unknown Population Variance)

When the population variance is not known, it is not appropriate to use the z-test, even when the population is normally distributed, because the σ is unknown.

$$Z = \frac{X - U}{\sigma/\sqrt{n}}$$

An appropriate test statistics becomes-

$$T = \frac{X - U}{s/\sqrt{n}} \quad \text{Where } S = \sqrt{\sum \frac{(X - \overline{X})^2}{n - 1}}$$

$$\text{or} = \sqrt{\sum \frac{(X - \overline{X})^2}{\sum f - 1}}$$

This statistic has a student t – distribution with (n – 1) degree of freedom.

Example 2

The mean operating temperature of a certain devise, according to the manufacturer, is greater than 190 degrees. A mean and standard deviation of 195 and 8 degrees, respectively are computed from the operating temperature of a random sample of 16 devices. Do these data provide sufficient evidence to indicate that the mean operating temperatures is higher than claimed?

Let $\alpha = 0.05$ and assume that operating temperature s are, approximately, normally distributed.

Solution

Step I: Hypotheses

$H_0 : \mu \leq 190$

$H_a : \mu \geq 190$

Step II: Significant level

$\alpha = 0.05$

Step III : t_{n-1} , $0.05 = t\,(16 - 1)\,0.05 = T_{15}, 0.05$

$= + 1.753$

Step IV: Test statistics

$$T = \frac{\overline{X} - U}{s/\sqrt{n}}$$

$$= \frac{195 - 190}{\frac{8}{\sqrt{16}}} = \frac{5 \times 4}{8}$$

$$T = \frac{5}{16} = 2.5$$

Since $T_{cal} > C_{tab}$ or critical- we reject H_{01} and conclude that the mean is greater than 190.

Since 2.131 < 2.5, P > 0.025 – hence, we reject H_0 conclude as above.

We can therefore, subsume that the necessary and sufficient conditions for the use of the t – distribution are:

1. the sample standard deviation, *s* is used to estimate the unknown population standard deviation σ.
2. the sample size is small (n < 30)
3. the population is approximately distributed.

3.13 The Differences between the Means of Two Normally Distributed Population

Here, you will learn about:

a. when the population variances are known
b. when the population variances are not unknown

Known population variances

The appropriate test statistics to use when the population variance are know and as usual, the population are normally distributed, is base on:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2 - U_1 - U_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}}$$

Where $\overline{X}_1 - \overline{X}_2$ is the difference between the sample means.

Example 3

Two procedures can be used to manufacture a certain product for which tensile strength is an important characteristic past experience has shown that the tensile strength resulting from both procedures are, approximately, normally distributed. A random sample of 12 items produced by procedure 1 gives a mean of 40ps, while a random of 16 items produced by procedure 2 yields a mean of 34ps. The standard deviation for procedure 1 is 6ps, and for procedure 2, the standard deviation is 8. Management wishes to know how the mean tensile of item produced by the two methods are different.
Let $\alpha = 0.05$.

Solution

Hypothesis-

H_0: $\mu_1 - \mu_2 \geq 0$ Ha: $\mu_2 < 0$
Significance level $\alpha = 0.05$
Critical $t_{n1\ tn2-1,\ 0.05} = t_{n0,}\ 0.05 = 1.717$

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_3}}}$$

$\overline{X} = 40, \quad X = 34$
$\sigma^2 = 6, \sigma_2^2 = 6 = 8, n_1 = 12, n_2 = 16$

$$Z = \frac{40 - 34 - 0}{\sqrt{\frac{\sigma^2}{12} + \frac{8^2}{16}}}$$

$$= \frac{6}{\sqrt{\frac{36}{12} + \frac{64}{16}}}.$$

$= 2.27$

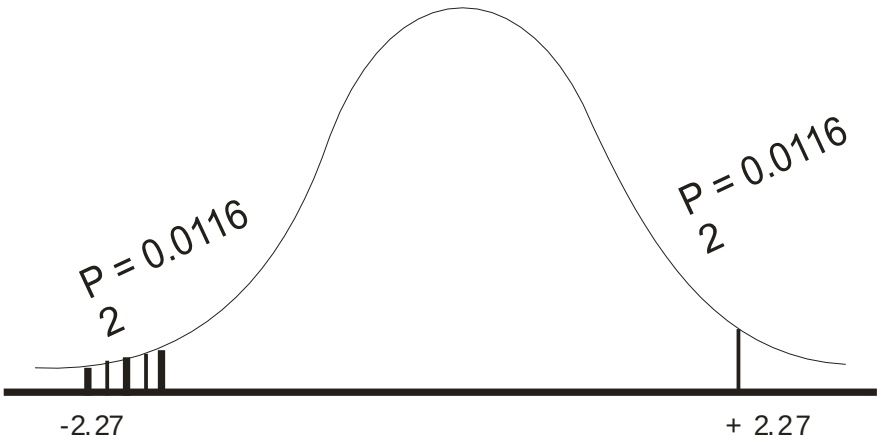


Fig. 1.3

0.5 – 0.4884 = 0.0116
Since it is two – tailed test,
P = 2 (0.0166).

Therefore, we conclude that the two items on the average do yield the same tensile strength.

Unknown population variances

Here, we shall treat this under two dichotomies

(i) Equal variance
(ii) Unequal variance

i. Equal population variance

When it has been established that populations are normally distributed and that they have equal population variance, the test statistics becomes:

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\bar{u}_1 - \bar{u}_2)}{\sqrt{\frac{S_P^2}{n_1} + \frac{S_P^2}{n_2}}}$$

Where-

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n - 1)S_2^2}{n_1 + n_2 - 2}$$

-the pooled estimate or the common population variance, with $n_1 + n_2 - 2_1$ degree of freedom.

Example 4

The table below shows the purchase of yarn from two vendors.

Table 1.1

	VENDOR 1	VENDOR 2
N	10	12
X	94	98
S^2	14	9

Based on an approximate hypothesis test with $\alpha = 0.05$, would you advise the textile manufacturer to buy less expensive yarn? Assume that the population variance is equal. Let $\alpha = 0.05$

Solution

Hypotheses-

$H_0 : \mu_1 = \mu_2 = 0$ $H_a - \mu_1 - \mu = < 0$
Significance level-
$A = 0.05$
Critical region t_{n1}, t_{n-2}, 0.05 = t_{20}, 0.05 = -1.717

$$T = \frac{(x_1 - x_2) - (u_1 - u_2)}{\sqrt{\frac{S_P^2}{n_1} + \frac{S_P^2}{n_2}}}$$

$$= \frac{94 - 98 - 0}{\sqrt{\frac{S_P^2}{10} + \frac{S_P^2}{12}}}$$

$$S_P^2 = \frac{(10 - 1)\,14 + 9\,(12 - 1)}{10 + 12 - 2}$$

$$= \frac{9 \times 14 + 9 \times 11}{20}$$

$$= \frac{126 + 99}{20} = 11.25$$

$$T = \frac{-4}{\sqrt{\frac{11.25}{10} + \frac{11.25}{12}}}$$

$$T = \frac{-4}{\sqrt{11.25 \pm 0.9375}}$$

$$= \frac{-4}{1.436}$$

$$= -2.79$$

Since we reject H_0
Since $T_{cal} < t_{critical}$, we reject H_0
- 2.845 < -2.79 < 2.525
- 0.01 > P0.05

Since 2.27 > 1.96, we reject H_0

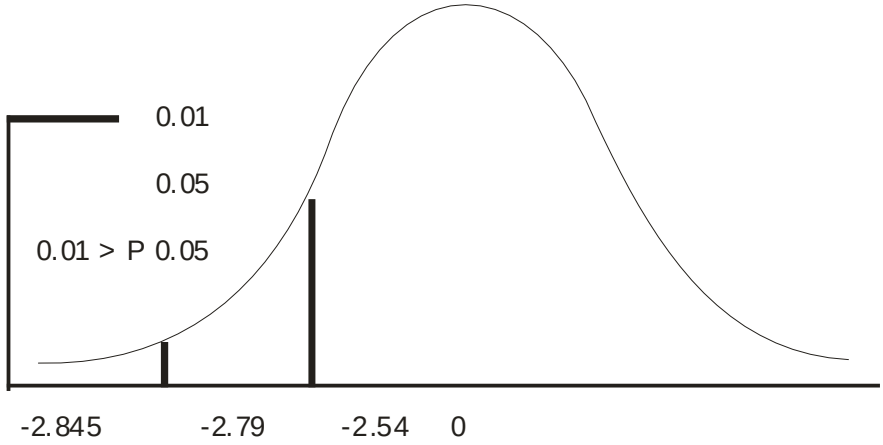


Fig. 1.4

We conclude that the textile material for vendor is lesser than that of vendor 2.

ii. Unequal variances

When the population variances are not equal, there is, of course, no basis for pooling σ^2 ad σ^2. The test statistics becomes:

$$T^{l} = \frac{(x_1 - x_2) - (u_1 - u_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Which distributed normally with student's degree of freedom.

$$df^{l} = \frac{\left[\left(\frac{S_1^2}{n_1}\right) + \frac{S_2^2}{n_2}\right]}{\sqrt{\frac{\left(\frac{S_1^2}{n_1}\right)}{n_1} + \frac{\left(\frac{S_2^2}{n_2}\right)}{n_2}}}$$

Example 5

Suppose in example 25 that we do not know whether the population variances are equal, and we are unwilling to assume that they are equal. Then-

$$t^2 = \frac{94 - 98 - 0}{\sqrt{\frac{14^2}{10} + \frac{9^2}{12}}}$$

$$= \frac{-4}{\sqrt{1.4 + 0.95}}$$

$$= \frac{-4}{\sqrt{2.15}}$$

$$= \frac{-4}{1.4663}$$

$$= -2.73$$

$$df^{l} = \frac{\left[\frac{14}{10} + \frac{9}{12}\right]^2}{\frac{\left(\frac{14}{10}\right)^2}{10} + \frac{\left(\frac{9}{12}\right)^2}{12}} = \frac{4.6225}{0.242875} = 19$$

Let $\alpha = 0.01$
Critical value = t_{19}, 0.01 = -2.539
Since -2.73 < - 2.539
We reject H_0
0.01 < P > 0.05

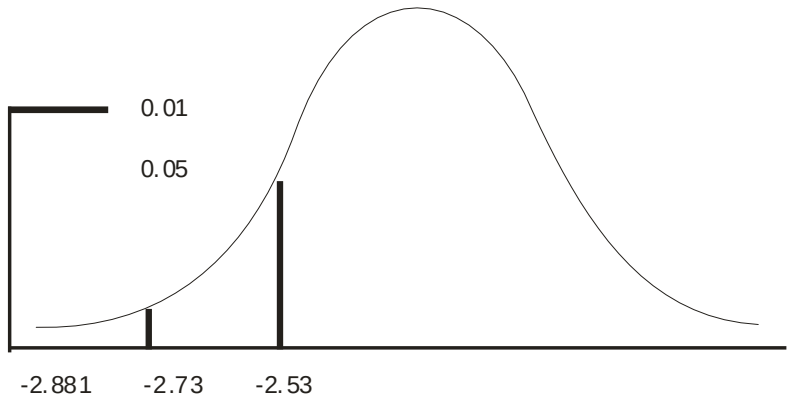


Fig.1.5

3.14 The Difference between the Means of Two Populations not Normally Distributed

The whole gamut of this assumption is that since the population is not normally distributed (abnormal distribution); we make use of the central limit theorem of the sample size-which is large. This enables us to use normal theory, since the sampling distribution will be approximately normally distributed.

The test statistic becomes:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (u_1 - u_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

Example 6

In study of the advertising practice of two type of retail firm, one variable of interest is the amount spent on advertising during the preceding year. An independent random sample is drawn from every other type of firm, with the following results. Can we conclude from these data that type A, firm spent more on advertising- on the average, than type did type B firm? $\alpha = 0.05$.

Solution

Type A	**Type B**
n : 60	70
$\bar{x}$i : N14,800	N14,500
s^2 : 180,000	133,000

Solution

The functional forms of the production are not given, however, the sample size is large; we rely on the central limit theorem.

Hypothesis-

$H_0: \mu_1 - \mu_2 \leq$ $H_a : \mu_1 - \mu_2 > 0$
$\alpha = 0.05$
Critical $Z_{0.05} = + 1.645$

$$t^{l} = \frac{(x_1 - x_2) - (u_1 - u_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$Z = \frac{14800 - 14500}{\sqrt{\frac{180{,}000}{60} + \frac{133{,}000}{60}}}$$

$$= \frac{300}{\sqrt{3000 + 1900}}$$

$$= \frac{300}{70}$$

$$= 4.29$$

Since $Z_{cal} > Z_{critical}$ we reject H_0.

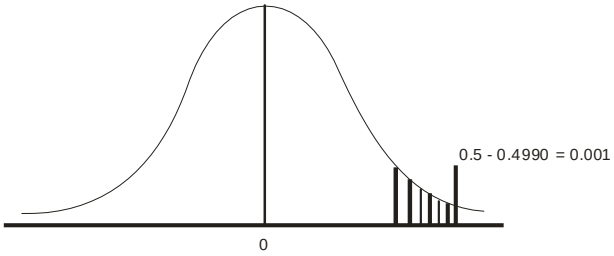


Fig. 1.6

Since 4.29 > 3.0 P < 0.001

Note that for a single population which is not normally or abnormally distributed, the central limit theorem holds, if the sample is large, for instance ($n \geq 30$). The test statistics becomes-

$$Z = \frac{(\bar{x} \quad u)}{\sigma\sqrt{n}}$$

3.15 Paired Observations

Suppose we want to compare two population and we want to do so by finding means difference μ d of this populations. The test statistics that is suitable is a paired comparison test.

In this case-
$d_i = x_i - x_{2i}$, where x_{ii} and x_{ii} are the observations taken on the *ith* pair or object under condition 1 and 2 respectively.

The test statistic, when the population is normally distributed and the population variance of the difference is indicated thus-

$$Z = \frac{\bar{d} - u_{do}}{\sqrt{\bar{d}}}$$

When the variance is unknown or the sample is too small (n<30) the test statistics is-

$$t = \frac{\bar{d} - u_{do}}{S\bar{d}}$$

Where d =

$$sd = \sqrt{\frac{\leqslant D^2 - (\leqslant d)^2}{n(n-1)}}$$

$$\text{And } s\bar{d} = \frac{sd}{\sqrt{n}}$$

Example 7

The times taken by seven workers to perform a task are measured in minutes, first before training and then after training. Test the hypothesis that:

(1) the training programme has made no difference to performance.

Let $\alpha = 0.05$

Table 1.2

WORKERS	**BEFORE TRAINING**	**AFTER TRAINING**
	X_1	X_2
1	7	8
2	6	6
3	2	10
4	5	12
5	4	9
6	6	10
7	8	9

Solution

Table 1.3

X_1	X_2	$d = (X_1 - X_2)$	d^2
7	8	-1	1
6	6	0	0
2	10	-8	64
5	12	-7	49
4	9	-5	25
6	10	-4	16
8	9	-1	1
		$\Sigma d = -26$	$\Sigma d^2 = 156$

$$= \sqrt{\frac{\Sigma D^2 - (\Sigma d)^2}{n\,(n-1)}}$$

$$= \sqrt{\frac{7 \times 156 - (-26)^2}{7(6)}}$$

$$= \sqrt{\frac{7 \times 156 - 676}{42}}$$

$$S_{\bar{d}} = \sqrt{\frac{1092 - 676}{42}} = 3.15$$

$$\bar{d} \quad \frac{\Sigma d}{n} = \frac{-26}{7} = -3.71$$

$$S_{\bar{d}} = \frac{3.15}{\sqrt{n}} = \frac{3.15}{\sqrt{7}} = 1.19$$

Hypothesis-

$H_0 = \mu_d = 0 \quad H_a: \mu_D = 0$
$\alpha = t\ 0.05, n - 1\ t\ \alpha/2 = 6.025 = \pm 2.447$

$$t = \left| \frac{\bar{d} - Ud0}{sd} \right|$$

$$= \left| \frac{\bar{3}.71 - 0}{1.19} \right|$$

$$= \left| \frac{-3.71}{1.19} \right|$$

$$= 3.12$$

Since a /t/ > /tα /2, i.e 3.12 > 1.19

We reject H_0 and conclude that the training programme makes a difference. Since 3.12 > 2.998 P < 0.01.

3.16 Testing a Hypothesis about a Population Proportion

The approximate test statistic for testing hypothesis about population proportion is:

$$Z = \frac{\bar{P} - \bar{P}}{\sqrt{\frac{p_o\ q_o}{n}}}$$

Where P_0 is the hypothesis proportion, $q_0 = 1 - P_0$ and P is the sample proportion.

Example 8

The president of a certain firm concerned about the safety record of the firm's employees, sets aside N15, 000.00- a year, for safety, education and promotion. The firm's accountant believes that more than 75% of similar firms spend more than N15,000.00 a year on safety education and promotion.

Let α = 0.05, do you agree with the accountant? If some of the 60 firms state that they spend more than N15, 000.00 per year on safety education and promotion.

Hypothesis-

H_0 $P \leq 0.075$ H_a : $P > 0.75$
$\alpha = 0.05$

Critical region $0.05 = \pm 1.645$

$$\bar{P} \frac{50}{60} = 0.83$$

$$Z = \frac{\bar{P} - \bar{P}}{\sqrt{\frac{p_o \, q_o}{n}}}$$

$$= \frac{0.83 - 0.75}{\sqrt{\frac{(0.75)\,(0.25)}{60}}}$$

$$= \frac{0.08}{0.0559} = 1.43$$

Since 1.43 < 1.645, we accept the null hypothesis.

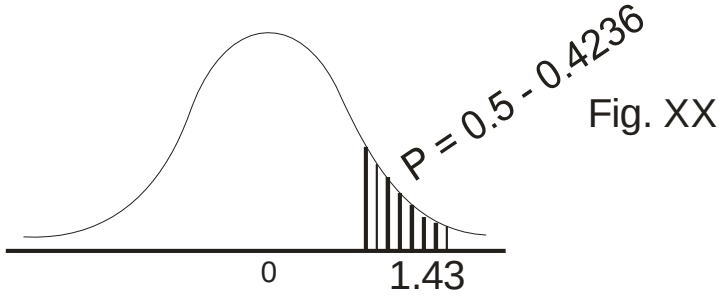


Fig. XX

P = 0.0764

Fig. 1.7

The proportion may be less than or equal to 0.75

3.17 Testing a Hypothesis about The Difference between Two Population Proportion

The test statistic is as follows-

$$Z = \frac{(\bar{P}_1 - \bar{P}_2) - (P_1 - P_2)}{\sqrt{\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}}}$$

Where $P_1 = P_2 = \frac{X_1 + X_2}{n_1 + n_1}$

$$\bar{P}_1 = \frac{X_1}{n_1 + n_2} \quad \bar{P}_2 = \frac{X_2}{n_1 + n_2}$$

In this instance, we are making use of the pooled estimate to computer the standard error.

$$S_{P_1 - P_2} = \sqrt{\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}}$$

$$Z = \frac{\bar{P}_1 - \bar{P}_2 - 0}{S_{P_1 - P_2}}$$

In the other hand, the unpooled estimate of the error is:

$$S_{P_1 - P_2} = \sqrt{\frac{P_1(1 - \bar{P}_1)}{n_1} + \frac{P_2(1 - \bar{P}_2)}{n_2}}$$

Example 9

The ability to withstand temperature of up to 250^0F is necessary requirement of a material used in the manufacture of a certain item. Two available materials- one a natural material, the other a systematic and more economical material- are equally satisfactory in all respects, except possibly, in the area of heat tolerance. Simple random samples of 225 specimen of each of the two materials are tested for these characteristics. The samples are independently drawn.

Thirty-six specimen of the natural material and 45 specimen of the synthetic material fail at temperature below 250^0F. Can we conclude from these data that the two materials are different with respect to heat tolerance?

Let $\alpha = 0.05$.

Solution

Hypothesis-

H_o: $P_1 - P_2 = 0$, $P_1 - P_2 \neq 0$
$\alpha = 0.05$, $\alpha/_2 = 0.025$
Critical region $0.025 = \pm 0.196$

$$\text{Where } P_1 = P_2 = \frac{X_1 + X_2}{n_1 + n_1} = \frac{36 + 45}{225 + 225} = 0.162$$

$$\bar{P}_1 = \frac{36}{225} = 0.16 \qquad P_2 = \frac{45}{225} = 0.2$$

$$Z = \frac{(\bar{P}_1 - \bar{P}_2 - 0}{\sqrt{\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}}}$$

$$= \frac{0.16 - 0.2}{\sqrt{(0.162)(0.638) + (0.162)(0.838)}}$$

$$= \frac{-0.04}{0.0347} = -1.153$$

Since, $-1.153 > -1.96$

We do not reject the null hypothesis-

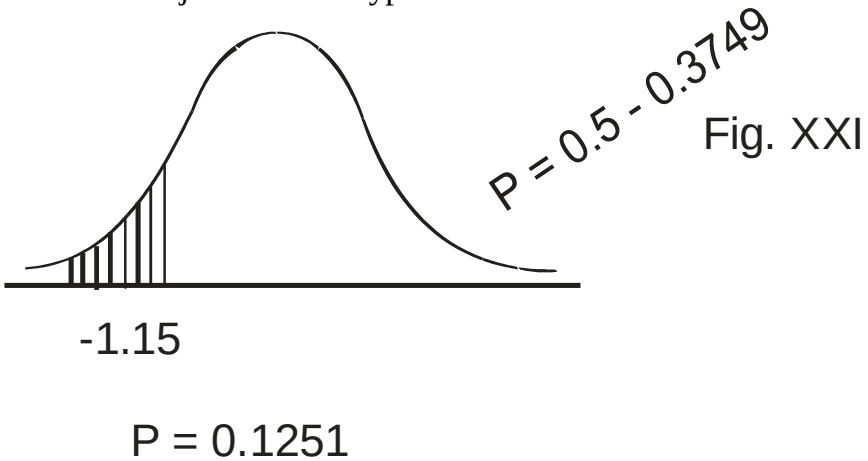


Fig. 1.8

Using the unpooled standard error, we have-

$$\text{Where } P_1 = P_2 = \frac{X_1 + X_2}{n_1 + n_1} = \frac{36 + 45}{225 + 225} = 0.162$$

$$\bar{P}_1 = \frac{36}{225} = 0.16 \qquad P_2 = \frac{45}{225} = 0.2$$

$$Z = \frac{(\bar{P}_1 - \bar{P}_2 - 0}{\sqrt{\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}}}$$

$$Z = \frac{0.16 - 0.2}{\sqrt{\frac{0.16(0.84)}{225} + \frac{0.2\,(0.8)}{225}}}$$

$$= \frac{0.04}{\sqrt{0.001308}}$$

$$= \frac{-0.04}{0.001308}$$

$$= 1.11$$

Therefore, we conclude that pooling has had a negligible effect.

3.18 Testing a Hypothesis about the Variance of a Normally Distributed Population

The approximate test statistics for testing-

H_0: $\sigma^2 = \sigma^0$ is

$$X^2 = \frac{(n - 1)s^2}{\sigma^2}$$

Where, S^2 is computed from a random sample of size *n* from a normally distributed population. When the null hypothesis is true, the test statistics is distributed as (chi-square) with n – 1 degree of freedom.

Example 10

A simple random sample of size 21 from a normally distributed population gives a variance of 10. Test the null hypothesis that $\sigma^2 = 1.5$ against the alternative $\sigma^2 = 1.5$

Let $\alpha = 0.05$

Solution

Hypothesis-

H_0: $\sigma^2 = 15$ $\sigma^2 \neq 15$
$\sigma = 0.05$
$\lambda^2 = \frac{(21 - 1)10}{15}$, $21 - 1 = 20$, $\frac{0.05}{2}$ (0.025)

Critical region 20, $\frac{0.05}{2}$ = 20, 0.025
$X^2 = \frac{20 \times 10}{15} = 13.33$
Critical region 20, 0.05 = 20, 0.025 = 34.1696

Since 13.33 < 34.1696, we can not reject H_0. Since 13.33 > 12.4426
P < 0.1

Note that the sample variance S^2 is calculated by

$$S^2 = \frac{\sum (X - \bar{X})^2}{n - 1} \text{ or } \frac{\sum (X - \bar{X})^2}{\sum f - 1}$$

3.19 The Ratio of the Variance of Two Normally Distributed Populations

In comparing two variances, we use the test statistics.

$F = \frac{S_1^2}{S_2^2}$, with Fi distribution test, that have,
$V_1 - 1$, and $V_2 - 1$ degrees of freedom where $V_1 - 1$ is numerator and $V_2 - 1$ denominator.

When a person was considering the use of that test to the difference between two means, two samples of size 16 yield variance of 28.5, and 9.5, respectively. Do the data indicate that the f –test is inapproximate

on the basis of the assumption of equality of population variance? Let $\alpha = 0.05$

Solution

Hypothesis-

$H_0: \sigma_1^2 = \sigma_2^2$
$H_a: \sigma_1^2 \neq \sigma_2^2$
$\alpha = 0.05$
$\alpha/_2 = 0.025$

Critical region $(\alpha/_2, (\alpha_2 - 1) (V_2 - 1)$

$= 0.025, 15, 15$
$= 2.86$
$F = \frac{S_1^2}{S_2^2}, \quad \frac{28.5}{9.5} = 3$
Since $2.86 < 3$, we reject H_0
$-+ \ 0.01 < P < 0.025$

SELF-ASSESSMENT EXERCISE

A researcher conducted a study of the grocery-shop habits of residents of a certain city; the study (via interview), covering about 400 households revealed the following information- of 225 shoppers with rural backgrounds and 175 shoppers with urban backgrounds, 54 and 52, respectively, state that they do most of their grocery shopping at a chain store. We want to decide on the basis of this sample, whether or not the two groups differ with respect to where they do most of their grocery shopping. Let $\alpha = 0.01$.

4.0 CONCLUSION

You have learnt how to test hypothesis in this unit. You now know the various steps to take when conducting hypothesis testing. This aspect of statistics is very crucial because you have to take decisions from time to time- in everyday life; you even need it in your research project work. You can now select the best test statistic in the course of verifying your hypothesised population parameters. You will be able to excel in inferential statistics as you apply all that you learned in this unit. You can now prepare yourself for the next unit on another aspect of inferential statistics- correlation and regression analysis.

5.0 SUMMARY

The following are the major highlights of this unit.

- Acceptance region- region in which the null hypothesis cannot be ejected.
- Alternative hypothesis- this specifies the hypothesis assumed true, if the null hypothesis is rejected.
- Critical region- region in which the null hypothesis cannot be accepted (sometimes referred to as ejection region).
- Critical value- a value that is compared with the test statistic to determine whether (H_o) is accepted or rejected.
- Hypothesis testing procedures using sample statistics to test hypothesised values of population parameters.
- Level of significance- the maximum probability of a type 1 error that the user will tolerate in the hypothesis testing procedure.
- Null hypothesis- this specifies the hypothesis value of the parameter to be tested.
- One- tailed test- an hypothesis test in which rejection of the null hypothesis occurs from values of the test statistic in one tail of the sampling distribution.
- P- value- the probability of observing a sample outcome even more extreme than the observed value when the null hypothesis is true.
- Rejection region- region in which the null hypothesis cannot be accepted (sometimes referred to as critical region).

6.0 TUTOR-MARKED ASSIGNMENT

i. A battery manufacturer claims that his batteries have an average life of 55 hours. A batch of 40 batteries is tested, given a mean life *x* of 50 hours with a standard deviation of 11.734 hours. Does this show that the manufacturer's claim is justified at the percent level of significance?

ii. A survey of a country's work force confirms that the number of days lost, each year, through sickness was 15 days per worker, on the average. Suppose that a researcher want to test this survey result by monitoring the records of a sample of 25 workers. The following data show the number of days of absence for each of the 25 workers during a particular year- 5, 25, 10, 0, 3, 50, 12, 14, 40, 12, 32, 8, 4, 47, 20, 14, 16, 10, 1, 22, 58, 5, 23, 18, 9. Use these data to determine whether the claim of loss of 15 days per worker, each year, should be rejected by he researcher's sample survey result (at the five percent level of significance).

7.0 REFERENCES/FURTHER READING

Brian, P.M. & Philip, M.N. (2006). *Applied Statistics for Public Policy*. New Delhi: Prentice-Hall.

James, T. & Benon, P.G. (1988). *Statistics for Business and Economics*. (4th ed.). United States: Dellen Publishing Company Inc.

Kasumu, R. B. (2002). *Introduction to Probability Theory*. *A First Course*. JAB Publishers.

UNIT 2 CORRELATION AND REGRESSION ANALYSIS

CONTENTS

1.0 INTRODUCTION

In the previous unit, you were exposed to statistical inferences from sample data, by testing hypothesis about population parameters. Now this unit will take you through the analysis of relationships between variables. The unit introduces simple and multiple regression and correlation analysis. For instance, in analysing data generated by a businesses or industrial operation, we are often interested in knowing

something of the relationship between two variables *X* and *Y*, called the independent variable and dependent variable.

In other words, we are interested in knowing the relationship between more than two variables; in which case, we have one dependent variable *Y* and, at least, two independent variables X. The former is called simple linear regression, while the later is referred to as multiple regressions. On the other hand, correlation analysis is concerned with measuring the strength of the relationship between variables. In this unit, we shall examine both. We shall also test the statistical significance of the regression and correlation models.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- state the usefulness of measuring relationship between variables
- explain regression results for estimation and prediction and create confidence and predictive intervals
- explain the regression plane
- interpret coefficient correlation and determination.

3.0 MAIN CONTENT

3.1 Simple Linear Regression Model

Suppose we are interested in studying the relationship between sales of a certain product and the age of persons in the various market areas. Here, we can identify two variables. We call the sales of certain product '*Y*' and the age of persons in various market an area, '*X*'. If there is a linear relationship, we can therefore express the linear model thus:

$$Y_i = \alpha + \beta x_i + e_{ij}$$

Where y_i is the value of the *y* variable for a typical unit of association from the population; X_i is the value of *X* variable for that same unit of association, α and β are parameters called the regression constant and regression coefficient, respectively and *ei* is a random variable with mean μ and a variance of σ^2.

3.2 Fitting a Simple Linear Regression (SLR) Model

As shown above, the *SRL* model is given by the relation-

$Y_i = \alpha + \beta X_i + e_{ij}$. This means we need to calculate the unknown parameters α (alpha) and β (Beta). Let $\hat{\alpha}$ (alphacap) and $\hat{\beta}$ Betacap) be the estimations of α and β respectively. To achieve this feat, there are two methods we can use, namely.

(a) Fitting simple linear regression by eye and
(b) Fitting by the method of least squares

3.3 Eye Fit Method

This method is based on scatter diagram, which is a graph of the observed pairs of observations. We assign the value of the independent variable *X* to the horizontal axis and assign the values of the dependent variable *Y* on the vertical axis. We place a dot on the graph at the intersection of each pair of values. Then, we draw the best straight line our eyes can give us:

(i) calculate $\overline{X}$ (mean of x)
(ii) calculate $\overline{Y}$ (mean of Y)
(iii) plot ($\overline{X}$, $\overline{Y}$) and draw the best straight line to pass through the intersection of X and Y reads (bar X and Y).
(iv) the intercept α (constant) slope (β) are obtain as:

β = slope of line = $\tan\theta = \frac{\text{opposite}}{\text{adjacent}}$
and α, point where it cuts the Y – axis.

Example 1

Draw a scatter diagram for the following data.

X	2	3	5	6	9	10	11	14
Y	3	2	5	8	7	9	11	12

Fit a regression line of *y* on *x* as best as you can. From your graph obtain the regression coefficients of *y* and *x*.

$$X = \frac{\Sigma x}{9} = 7.6 \quad Y = \frac{61}{9} = 6.8$$

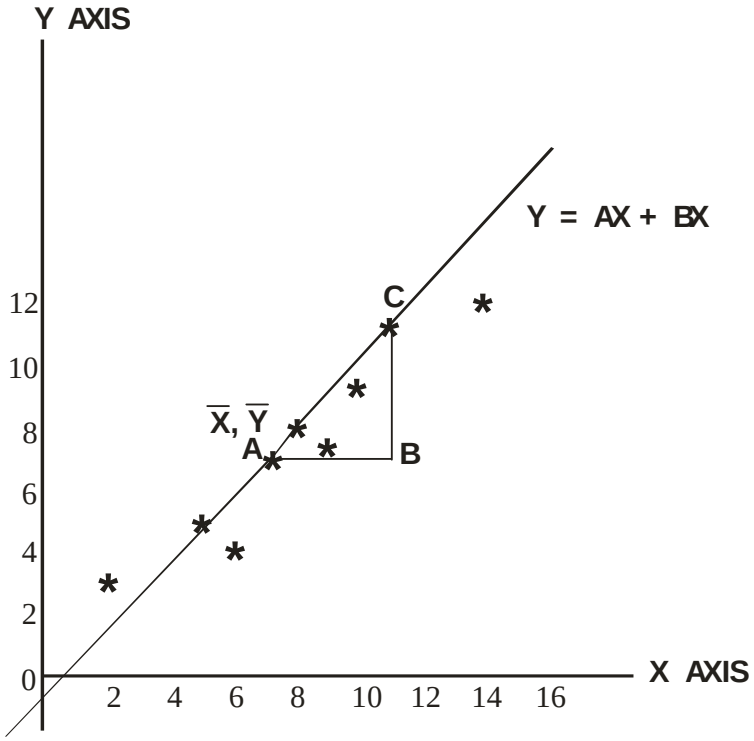


Figure 2.1: Eye Fitting of Regression Line

$$a = \beta = \frac{opp}{adj} = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{11 - 6.8}{11 - 7.6} = \frac{4.2}{3.4} = 1.2$$

$b = \alpha = -2.3$

$\hat{Y} = 1.2x - 2.3$

3.4 Least Squares Method

The eye fitting method is myopic and very inaccurate and unreliable. The most accurate and reliable device is the least squares method. We derive it thus;

$$Y_i = \alpha + \beta X_i + e_i$$
$$e_i = Y^i - \alpha - \beta Xi$$

by minimising the sum of squares of *ei* (errors) we have-

$$\Sigma e_i^2 = (yi - \alpha - \beta Xi)^2.$$

Using our knowledge of calculus, by differentiating partially with respect to α and keep β constant to form equation (1) and in turn differentiating partially with respect to α, keeping β constant to form equation (2) and solve simultaneously thus-

$$\frac{\sigma \Sigma e_i^2}{\sigma \alpha} = \Sigma (yi - \alpha - \beta X_i)^2 = 0$$

$$\Sigma y_i - n\alpha - \beta \Sigma Xi = 0$$

$\Sigma y_i = n\alpha + \beta Xi$ _____(1)

$\frac{\sigma \Sigma e_i^2}{\sigma\beta} = 2\Sigma (yi - \alpha - \beta X_i) X = 0$

$\Sigma X_i yi = \alpha \Sigma X + \beta \Sigma X_i$ --------(2)
1 x by ΣX: $\Sigma X_i \Sigma Y_i = \alpha n \Sigma X_i + \beta(\Sigma X_i)^2$ --------- (3)
2 x n : $n \Sigma X_i \Sigma Y_i = \alpha n \Sigma X_i + n \beta \Sigma X_i^2$ --------- (4)

Eqn 2 – 1: $n \Sigma X_i \Sigma Y_i - \Sigma X_i \Sigma Y_i = n \beta \Sigma X_i^2 - \beta(\Sigma X_i)^2$

$$\beta = \frac{n \Sigma X_i \Sigma Y_i - \Sigma X_i \Sigma Y_i}{\beta \Sigma X_i^2 - \beta(\Sigma X_i)^2}$$

From equation 1-

$$\Sigma y_i = n\alpha - \beta \Sigma Xi$$
$$\Sigma y_i + \beta \Sigma Xi = n\alpha$$
$$\alpha = \frac{\Sigma y_i}{n} - \frac{\beta \Sigma Xi}{n}$$
$$\alpha = \bar{Y} - \beta \bar{X}$$

Where $\bar{Y}$ is mean of *Y* and $\bar{X}$ is mean of *X*.

Example 2

Compute the least squares method; calculate the predicted values (b) the residuals. (c) the values of *Y,* when x = 4 and 15.

Solution

Table 2.1

n	X	-y	XY	X^2	Predicted $\hat{Y} = 0.188+0.87x$	Residuals $(y- \hat{Y})$	Y^2
1	2	3	6	4	1.928	1.072	9
2	3	2	6	9	2.798	-0.798	4
3	5	5	25	25	4.538	0.462	25
4	6	4	24	36	5.408	-1.408	16
5	8	8	64	64	7.148	0.852	64
6	9	7	63	81	8.018	-1.018	49
7	10	9	90	100	8.888	0.112	81
8	11	11	121	121	9.758	1.242	121
9	14	12	168	196	12.368	-0.368	144
n=9	$\Sigma x=68$	$\Sigma y=61$	$\Sigma xy=567$	$\Sigma x^{2=636}$		$\Sigma ei = 0$	$\Sigma y^2 = 513$

$\bar{X} = \frac{68}{9} = 7.6$

$\bar{Y} = \frac{61}{9} = 6.8$

$\beta = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{} = \frac{9 \times 567 - (68)(61)}{} = 0.87$

$$\alpha = \frac{n\Sigma x^2 - (\Sigma x)^2}{} \qquad 9 \times 636 - (68)^2$$

$$\alpha = \bar{Y} - \beta \bar{X}$$

$$= 6.8 - 0.87\ (7.6)$$

$$\alpha = 0.188$$

From the result, you will observe why we reasoned above that the eye fitting method is unreliable and crude.

The regression line can be expressed as-

$$\hat{Y} = 0.188 + 0.877X \_\_\_\_\_ ****$$

The equation **** can be used to draw the best fitting regression line $\hat{Y} = \alpha + \beta X$.

(b) The predicted values of $\hat{Y} = 0.188 + 0.87X$ and the residual are shown in table

(c) When $X = 4$

$\hat{Y} = 0.188 + 0.87\ (4)$

$= 3.292$

When $X = 15$

$\hat{Y} = 13.238$

3.5 Testing the Statistical Significance of the Regression Model

A significance test is required to test the probability of obtaining such positive or negative (i.e. non-zero) sample coefficients even though the population parameters are really zero.

Therefore, we test the null hypothesis.

Ho: $\beta = 0$ and Ha: $\beta \neq 0$. Accepting that Ho: $\beta = 0$ means that the slope of the population regression line is horizontal line so that the value of *Y* does not vary as *X* varies. In this case, information about *X* would be of no value in helping us to predict value of *Y*. On the other hand, if $\beta \pm 0$ or $\beta > 0$ or $\beta < o$ (alternatives hypothesis) then the population regression line must slope upwards (or down wards) so that information about *X* will help us to predict values of y.

The test statistic *t* is used in regression analysis because appropriate population information is not available, and hence, a ***Z*** test cannot be used. The test statistic *t* is therefore given by-

$$t = \frac{\beta + \beta}{S_b}$$

Where S_b (standard error of b) $= \frac{SEE}{\sqrt{\Sigma x^2_i - n\overline{x}^2}}$

Where SEE (Standard Error Estimate) is computed as-

$$SEE = \sqrt{\frac{\Sigma e^2_i}{n-2}} \qquad \sqrt{\frac{\Sigma x^2_i - \alpha\Sigma yi - \beta\Sigma x_i yi}{n-2}}$$

With n – 2 degrees of freedom (when *n* is the number of paired x, y observations). The test statistic for the constant or intercept α -

$$t = \frac{\hat{a} - \alpha}{Sa}$$

Where $Sa = \frac{SEE\ (\Sigma x^2_i)}{n(\Sigma x^2_i - n\overline{x}^2)}$

Example 3

Using an earlier example, test the null hypotheses that-
Ho : β = 0, th: β±0 and Ho: α = 0, Ha α±0.
Let α = 0.05

Solution

Ho: β = 0, Ha : β±0

$$t = \frac{\beta - \beta}{S_b}$$

b = 0.87

β = 0

$$S_b = \frac{SEE}{\sqrt{\Sigma x^2_i - n\overline{x}^2}}$$

$$SEE = \sqrt{\frac{513 - 0.188(61) - 0.87(567)}{9-2}}$$

$$= \sqrt{\frac{513 - 11.468 - 493.29}{7}}$$

$$= \sqrt{1.7743}$$

$$= 10.778$$

$$\sqrt{\Sigma x_i^2 - nx^2} = \sqrt{636 - 9\ x7.6^2}$$

$$= \sqrt{636 - 519.84}$$

$$= = 10.778 \sqrt{116.16}$$

$$:. \quad t = \frac{0.87 - 0}{\frac{1.085}{40.778}}$$

$$t = \frac{0.87 \times 10.778}{1.085}$$

$$t = 8.64$$

$t \quad 0.^{05}/_{21} \quad n - 2 = \text{to } .025, 7 = 2.365$

Since, *tcal* > table we reject *Ho*, and conclude that β is not zero that there is a linear relationship between *X* and *Y*. since 8.64 > 3.499, P < 2 (0.005) = 0.01

Ha: $\alpha = 0$ Ha: $\alpha \pm 0$.

$$t = \frac{0.188.0}{SEE} = \frac{0.188}{\sqrt{\frac{\Sigma x^2_i}{n(\Sigma x^2_i - nx^2)}}} = \frac{1.085}{\frac{0.188}{\sqrt{\frac{636}{9(636 - 519.84}}}}}$$

$$\frac{0.188}{0.8463} \quad \sqrt{\frac{636}{10085}} \quad 10\ +045.44$$

$$= \mathbf{\underline{0.\ 222}}$$

Since *tcal* < table , we accept *Ho*, and conclude that α is equal to zero.

3.6 Estimation and Prediction Using The Regression Model

Given an earlier result, the regression equation (model) can be used, of course, to predict values of the dependent variable *Y* for any give values of *X*. These intervals are referred to, respectively, as-

(1) confidence intervals
(2) predictive intervals.

3.7 Confidence Interval

To construct a 100 (1 – α) percent confidence interval of $\hat{Y}$ for a given value of *X* (demoted *Xo,* we use the following formula-
Confidence interval

$$\hat{Y} \pm \alpha/_2 \quad SEE \sqrt{{}^{1}/_{n} + \frac{(Xo-X)^2}{\Sigma x^2_i - nx^2}}$$

Where $\hat{Y} = \alpha + \beta Xo$ and $t\ \alpha/_2$ has n – 2 degrees of freedom.

3.8 Prediction Interval

To construct *100 (1 – α)* percent prediction interval for the estimated individual value of *y(i.e* $\hat{Y}$*)* for given value of *x* (i.e Xo), the formula is the same as that above except that *"1"* is added to the expression within the square root sign:

Prediction interval

$$\hat{Y} \pm t\ \alpha/_2\ [\ SEE \sqrt{1 + {}^{1}/_{n} + \frac{(Xo-X)^2}{\Sigma x^2_i - nx^2}}$$

Where $\hat{Y} = a + b\ x_o$ and $t\ \alpha/_2$ has n – 2 degrees of freedom.

Example 4

It is thought that the number of industrial down times that occur each year is related to the level of unemployment. The data given below were collated over a ten year period.

(a) What would you regard as the dependent and independent variables?
(b) Construct 95 percent confidence intervals for an individual value of *S* when *U* = 3 *U* = 15.
(c) Construct 95% prediction intervals for an individual value *S* when *U* = 2 and year.

Table 2.2

Year	1	2	3	4	5	6	7	8	9	10
Stop(s)	2.5	2.9	2.9	2.3	2.0	2.7	2.5	2.1	1.3	1.3
Unemployment	3.8	2.7	2.6	4.0	5.5	5.8	5.7	5.3	6.8	10.5

Solution

(a) Independent variable U. Dependent: S

(i) Confidence intervals

$\hat{Y} \pm t\,\alpha/_2, n-2$ [SEE $\sqrt{1/_n + \frac{(Uo - \bar{U})^2}{\Sigma U^2_i - n\bar{u}^2}}$

Where $\hat{Y} = \alpha + \beta X_0$

$\hat{S} = \alpha + \beta U_0$

Let U = X and S = Y

$$\beta = \frac{n\,\Sigma US - (\Sigma U)(\Sigma S)}{n\Sigma u^2 - (\Sigma u)^2} = \frac{10(108.6) - (22.5)(52.7)}{10(325.45) - (52.7)^2}$$

$$= -0.209$$

$$\alpha = \Sigma\, {}^{s}/_{n} - \frac{\beta \bar{U}}{n}$$

$$= 2.25 + 1.10143$$

$$\alpha = 3.35$$

$$\hat{S} = \alpha + \beta \bar{U}$$

$$\hat{S} = 3.35 - 0.209U_0$$

When U = 3

$$\hat{S} = 3.35 - 0.209(3)$$

$$\hat{S} = 2.723$$

When U = 15

$$\hat{S} = 3.35 - 0.209\,(15)$$

$$= 0.215$$

$$SEE = \sqrt{\frac{\Sigma S^2_i - \alpha\,\Sigma S_i - \beta\,\Sigma U_i S_i}{10-2}}$$

$$= \sqrt{\frac{53.69 - 351\,(22.5) - (-0.209)\,(108.6)}{10-2}}$$

$$= 0.3557$$

$$\Sigma U^2_i - n\bar{u}^2 = 325.45 - (10)\,\frac{(52.7)^2}{10}$$

$$= 47.721$$

(b) Confidence intervals

$\hat{S} \pm t\,\alpha/_2, n-2$ [SEE $\sqrt{1/_n + \frac{(U_b - \bar{U})^2}{\Sigma U^2_i - n\bar{u}^2}}$

When U = 3, $t\,\alpha/_2, n-2$ t 0.025, 8 = 2.306

$$= 2.723 \pm (2.306)\,[(0.3557)\sqrt{1/_{10} + \frac{(3-5.27)^2}{7.721}}$$

$$= \quad 2.723 \pm (2.306)(0.3557)\sqrt{0.1 + 0.1080}$$

$$= \quad 2.723 \pm 0.3741$$

$$= \quad (2.349, 3.097)$$

When U = 5

$$\hat{S} = 0.215 \pm 2.306\ (0.3557)\sqrt{0.1 + \frac{(15-5.27)^2}{47.721}}$$

$$= 0.215 \pm 3.306\ (0.3557)\ (1.444)$$

$$= 0.215 \pm 1.698 = (-1.483, 1.913)$$

(c) Prediction intervals

$$\hat{S} \pm t\,\alpha/_2, n-2\ [\text{SEE}\sqrt{1 + {}^1/_n + (U_0 - \bar{U})}$$

$$\Sigma U^2_i - n\bar{u}^2$$

When U = 3

$$= 2.723 \pm 3.306\ (0.3557\sqrt{0.1 + 1 + 0.1080}$$

$$= 2.723 \pm 1.292$$

$$= (1.431, 4.015)$$

When U = 15

$$= 0.215 \pm 3.306\ (0.3557\sqrt{0.1 + 1 + 1.984}$$

$$= 0.215 \pm 3.306\ (0.3557)\ (1.7561)$$

$$= 0.215 \pm 2.065$$

$$= (-1.85, 2.28)$$

3.9 Extrapolation and Interpolation

Naturally, the regression equation may be used to forecast the values of the dependent variable *(y)* given the independent variable *(x)*. However, there is great danger in using the equation to forecast the values of $\hat{Y}$, outside the range of the observed data set from which the regression equation itself was derived. Computing values of the dependent variable *($\hat{Y}$)* within the range of recorded observations on the dependent variable *X* is referred to as **interpolation**. In contrast, projecting values (backwards as well as forwards) outside this range, given rise to $\hat{Y}$ values (Cap *Y* forecast), is referred as **extrapolation**.

From our previous example, we have forecasted that when U = 15, S = 0.215. This value falls outside the range of observed industrial stoppage values given in the question above. This means we are extrapolating. This value is far beyond the range of the previously observed values. Let say, we want to forecast the independent variable $\hat{Y}$, when X = 1, the forecast gives 3.141, the diagram below explains the result.

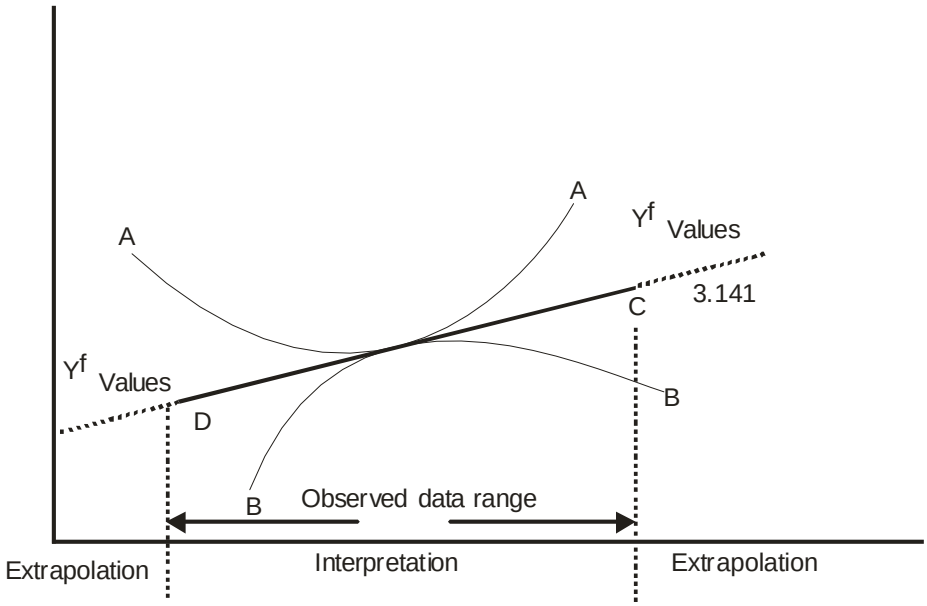


Fig. 2.2

Values within the range *CD* give rise to interpolation; while values that fall downwards or forwards in relation to *CD*, gives rise to extrapolation- e.g. 0.215 and 3.141. If we construct the upper and lower limits for various confidence and prediction intervals around a regression line, it will be seen that the width of the intervals increases, markedly, as the value of *U* in which we are interested (U_0) moves further away from the mean value of *U* (i.e. $\bar{U}$). This is shown in the figure below.

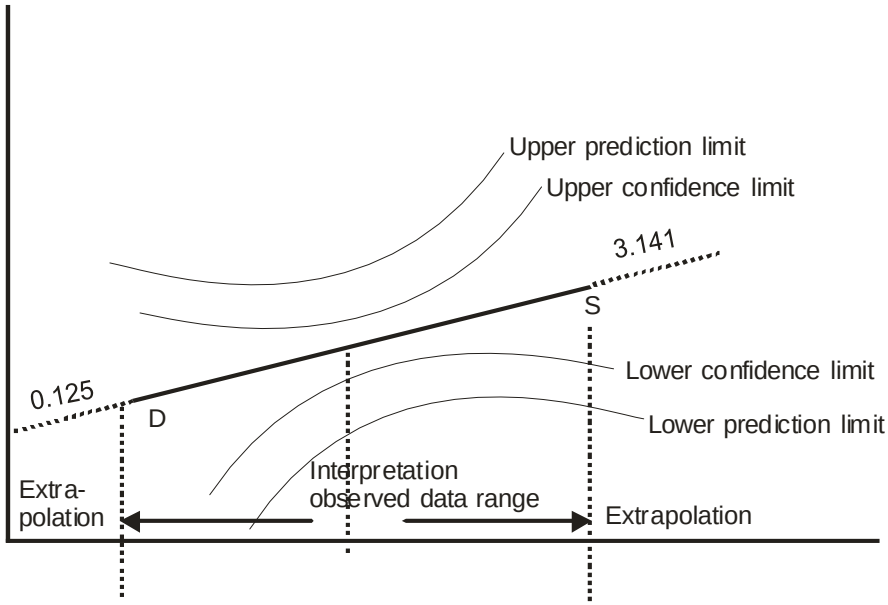


Fig. 2.3

From the figure, forecasts around extrapolation regions, therefore, are inevitably subject to wider margins of error. It will also be seen from this figure that prediction interval is wider than the confidence interval.

3.10 Multiple Regression

Simple linear regression involves two variables, one dependent variable *(Y)* and the other independent variable *X*, as we have mentioned earlier on. In multiple regressions, we have more than two independent variables. For example, the monthly sales of fuel (in thousands of litres) on average price *(f)* change per litre in each month and the advertising expenditure (N000) per month constitutes multiple regressions.

The linear model is given as-

$$Y = U0 + U_1X_1 + U_2 X_2 + \ldots\ldots\ldots + U_k X_k + eij$$

In this unit, we shall limit ourselves to three variables. One dependent variable and two independent variables:

$$Y = U0 + U_1X_1 + U_2 X_2 + \ldots\ldots\ldots + U_k X_k + eij$$

By transformation

$$eij = Y - U_0 - U_1X_1 - U_2U_2$$
$$eij^2 = (Y - U_0 - U_1X_1 - U_2U_2)^2$$
$$\Sigma e^2ij = \Sigma(Y - U_0 - U_1X_1 - U_2U_2)^2$$

By minimising the sum of squares of error and equate to zero, we shall differentiate, partially, with respect to U_0, U_1 and U_2.

i.e. $\frac{\partial \Sigma e^2ij}{\partial U_0} = 2\Sigma(Y - U_0 - U_1X_1 - U_2U_2) = 0$

$$\Sigma Y = nu_0 + u_1 \Sigma X_1 + U_2 \Sigma X_2 \quad \_\_\_\_\_(1)$$

$\frac{\partial \Sigma e^2ij}{\partial U_1} = 2\Sigma(X_1Y - U_0 - U_1X_1^2 - U_2 X_1U_2) = 0$

$$\Sigma X_1Y = u_0 \Sigma X_1 + U(\Sigma X_1^2 + U^2 \Sigma X_1X_2 \quad (2)$$

$\frac{\partial \Sigma e^2ij}{\partial U2_1} = -2\Sigma(X_1Y - U_0 - U_1X_1^2 - U_2 X_1U_2) = 0$

$$= \Sigma X_1Y - U_0 \Sigma X_2 + U_1 \Sigma X_1X_2 + U \Sigma X_2^2$$

We shall introduce matrices so that we can solve for U_0, U_1 and U_2. See mat 103. Using crammer's rule thus-

$$\begin{pmatrix} n & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X^2_1 & \Sigma X_1 X_2 \\ \Sigma X_1 Y & & \\ \Sigma X_2 & \Sigma X_1 X_2 & \Sigma X_2^2 \\ \Sigma X_2 Y & & \end{pmatrix} \begin{pmatrix} U_0 \\ U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} \Sigma Y \\ \\ \\ \end{pmatrix}$$

A X M

A is the coefficient matrix; X is the vector matrix and M is constant matrix

$$U_0 = \frac{/ DU_0 /}{/ A /}$$

$$U_1 = \frac{/ DU_1 /}{/ A /}$$

$$U_2 = \frac{/ DU_2 /}{/ A /}$$

$$/ DU_0 / = \begin{vmatrix} \Sigma y & \Sigma X_1 & \Sigma X_2 \\ \Sigma_1 Y & \Sigma X_1^2 & \Sigma X_1 X_2 \\ \Sigma X_2 Y & \Sigma X_1 X_2 & \Sigma X_2^2 \end{vmatrix}$$

$$/ DU_1 / = \begin{vmatrix} n & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1 Y & \Sigma X_1 X_2 \\ \Sigma X_2 & \Sigma X_2 Y & \Sigma X_2^2 \end{vmatrix}$$

$$/ DU_2 / = \begin{vmatrix} n & \Sigma X_1 & \Sigma Y \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1 Y \\ \Sigma X_2 & \Sigma X_2 Y & \Sigma X_2 Y \end{vmatrix}$$

Therefore-

$$U_0 = \frac{\begin{vmatrix} \Sigma y & \Sigma X_1 & \Sigma X_2 \\ \Sigma_1 Y & \Sigma X_1^2 & \Sigma X_1 X_2 \\ \Sigma X_2 Y & \Sigma X_1 X_2 & \Sigma X_2^2 \end{vmatrix}}{\begin{vmatrix} n & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1 Y & \Sigma X_1 X_2 \\ \Sigma X_2 & \Sigma X_2 Y & \Sigma X_2^2 \end{vmatrix}}.$$

$$U_1 = \frac{\begin{vmatrix} n & \Sigma Y & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1 Y & \Sigma X_1 X_2 \\ \Sigma X_2 & \Sigma X_2 Y & \Sigma X_2^2 \end{vmatrix}}{\begin{vmatrix} n & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1 X_2 \\ \Sigma X_2 & \Sigma X_2 X_2 & \Sigma X_2^2 \end{vmatrix}}$$

$$U_2 = \begin{vmatrix} n & \Sigma X_1 & \Sigma Y \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1 Y \\ \Sigma X_2 & \Sigma X_1 X_2 & \Sigma X_2 Y \end{vmatrix}$$

$$n \begin{vmatrix} \Sigma X_1 & \Sigma X_2 & \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1 X_2 \\ \Sigma X_2 & \Sigma X_1 X_2 & \Sigma X_2^2 \end{vmatrix}$$

Example 5

The following table shows the scores made by 10 assembly-line employees on a test designed to measure job satisfaction. The scores made on an aptitude test and the number of days absent during the past year (excluding vacations) is also shown. All employees were on the pay row during the entire year.

Table 2.3

Job satisfaction	**Aptitude(X_1)**	**Day absent (X_2)**
70	6	1
60	6	2
80	8	1
50	5	8
55	6	9
85	9	0
75	8	1
70	6	1
72	7	1
64	6	2

Find the multiple regression equation. Hence, find the job satisfaction when aptitude test is 4, and day absent is 3.

Solution

Using the method enumerated above, rows table takes this format

Table 2.4

n	**Y**	**X_1**	**X_2**	**X_1Y**	**X_2Y**	**X_1X_2**	**Y_1^2**	**X_2^2**	**Y^2**
1	70	6	1	720	70	6	36	1	4900
2	60	6	2	360	120	12	36	4	3600
3	80	8	1	640	80	8	64	1	6400
4	50	5	8	250	400	40	25	64	2500
5	55	6	9	330	495	54	36	81	3025
6	85	9	0	765	0	0	81	0	7225
7	75	8	1	600	75	8	64	1	5625
8	70	6	1	420	70	6	36	1	4900
9	72	9	1	504	72	7	49	1	5184
10	64	7	2	384	128	12	36	4	4096
10	ΣY = 681	Σx_1 =67	Σx_2 =26	$\Sigma x_1 y$ =4673	$\Sigma x_2 u$ =1510	$\Sigma x_1 x_2$ =153	Σx_1^2 = 463	ΣX_2^2 = 158	Σy^2 =47455

Model

$Y = U_0 + U_1 X_1 + U_2 X_2$

$U_0 = \frac{/ DU_0 /}{/ A /}$

$$U0 = \frac{\begin{vmatrix} 681 & 67 & 26 \\ 4673 & 463 & 153 \\ 1510 & 153 & 158 \end{vmatrix}}{\begin{vmatrix} 10 & 67 & 26 \\ 67 & 463 & 153 \\ 26 & 153 & 158 \end{vmatrix}} = \frac{298791}{8252} = 36.2$$

$$U1 = \frac{\begin{vmatrix} 10 & 681 & 26 \\ 67 & 4673 & 153 \\ 26 & 1510 & 158 \end{vmatrix}}{8252} = 5.39$$

$$U_2 = \frac{\begin{pmatrix} 10 & 67 & 681 \\ 67 & 463 & 4673 \\ 26 & 153 & 1510 \end{pmatrix}}{8252} = -1.62$$

$Y = 36.2 + 5.39X_1 - 1.62X_2$

When $X_1 = 4$ and $X_2 = 3$.

$Y = 36.2 + 5.39 (4) – 1.62(3)$

$= 36.2 + 21.56 - 486$

$Y = 52.9$

The method above is a bit time consuming and may give you stress. Here, is another method which we hope will reduce your stress. We shall use the transformations.

$b_0 = \hat{y} - b_1 x_1 - b_2 x_2$, where $\bar{Y} = \frac{\Sigma Y}{n}$, $\bar{x}_1 = \frac{\Sigma x_1}{n}$ and $\bar{x}_2 = \frac{\Sigma x_2}{n}$

where-

$b_1 \Sigma x^1{}_1{}^2 + b2\Sigma x^1{}_1 X^1{}_2 = \Sigma x_1 Y^1$ ******

$b1 \Sigma x^1{}_1 X^1{}_2 + B_2 \Sigma x^1{}_2{}^2 = \Sigma x_2{}^1 Y^1$ *****

$\Sigma x_1{}^2 = \Sigma x^1{}_1{}^2 - \frac{(\Sigma x_1)^2}{}$

$$= 463 - \frac{(67)^2}{10} = 14.1$$

$$\Sigma x_1^{1}{}_2^{2} = \Sigma x_1^{2} - \frac{(\Sigma x_1)^2}{n}$$

$$= 158 - \frac{(26)^2}{10}$$

$$= 90.4$$

$$\Sigma x_1 X_2^{1} = \Sigma x_1 X_2 - \frac{(\Sigma x_1)(\Sigma x_2)}{n}$$

$$= 153 - \frac{67 \times 26}{10}$$

$$= 153 - 174.2$$

$$= -21.2$$

$$\Sigma x_1^{1} y^{1} = \Sigma x_1 Y - \frac{(\Sigma x_1)(\Sigma y)}{n}$$

$$= 4673 - \frac{67 \times 681}{10}$$

$$= 110.3$$

$$\Sigma x_2^{1} Y^{1} = \Sigma x_2 y - \frac{(\Sigma x_2)\ \Sigma y)}{n}$$

$$= 1510 - \frac{681 \times 26}{10}$$

$$= -260.6$$

By substitution, we shall have :

$14.\ 1b_1 - 21.2b_2 = 110.3$ _____(i)
$-21.2b_1 + 2b_1 + 90.4b2 = -260.6$ _____(ii)

Solving this simultaneously we have –

$b_1 = 5.39$ b2 = -1.62. but
$b_0 = \hat{y} - b_1x_2 - b_2 x_2$
$= 68.1 - (5.39)(6.7) - (-1.62)(2.6)$
$= 68.1 - 36.113 + 4.212$
$= 36.2$
$Y = b_0 + b_1 X_1 + b_2 X_2$
$Y = 36.2 + 5.39 X_1 - 1.62X_2$

3.11 References about Individual Partial Regression Coefficients

We can make inferences regarding individual population partial regression coefficients when the assumptions of normality and equal

variances hold. In this case, μi is ,normally, distributed with mean β_1 and variance $Cii\ \sigma^2$; where *Cii* is found using Gauss multiplier since σ^2, the population variance in *Y*, is usually unknown, we use its estimate $S^2y.12…..k$.

The test statistic becomes-

$$t = \frac{b_1 - \beta_{10}}{Sb_i}$$

Where, Sbi = Sy. 12….k $\sqrt{Cii}$, with n – k – 1 degrees of freedom, using

t - destruction.

We may obtain the values of *C*, using Gauss multiplier.

$$C_{11}\Sigma x^1{}_1{}^2 + C_{12}\Sigma x_1 X_2{}^1 = 1$$
$$C_{11}\Sigma x_1{}^1 X_2 + C12\ \Sigma x^1{}_2{}^2 = 0$$
$$C_{12}\Sigma x^1{}_1{}^2 + C_{22}\Sigma x^1 X_2{}^1 = 0$$
$$C_{21}\Sigma x_1 X_2{}^1 + C_{22}\ \Sigma x^1{}_2{}^2 = 1$$

Where, $C_{12} = C_{21}$

Sy. 12 …..k = $\sqrt{MSE}$ = $\sqrt{\frac{SSE}{n-k-1}}$

SSE = TSS - RSS (Total sum of squares – Regression sum of square)

TSS = $\Sigma (Y_i - \bar{Y})^2 = \Sigma y^2{}_i - \frac{(\Sigma \bar{Y})^2}{n}$

RSS = $\Sigma (\hat{y}_i - \bar{Y})^2 = b_1\ \Sigma x_1{}^1 y + b_2\ \Sigma x_2{}^1 + 00.b_2\ \Sigma^1 k_i\ y_i$

Substituting into 7*S, we have

$14.1C_{11} - 21.2C_{12} = 1$ _____(1)
$- 21.\ 2C_{11} + 90.4C_{12} = 0$ _____(2)
$14\ .\ 1C_{21} - 21.2\ C_{22} = 0$ _____(3) *********
$-21.2C_{21} + 90\ .\ 4C_{22} = 1$ _____(4)
$C12 = C21$

Solving this simultaneously-

C_{11} = 0.0308
C_{12} = C_{21} = 0.0267
C_{22} = 0.0173.
TSS = $\Sigma Y^2 - \frac{(\Sigma \hat{y})2}{n}$ (from the table)
= $47455 - \frac{(68.1)^2}{10}$
= 47455 – 463.761

= 46991.24

RSS = $b_1 \Sigma x_{11} y^1 + b2 \Sigma x_2{}^1 y^1$

= 5.39 (110.3) + (- 1.62) (-260.6)

= 594.517 + 422.172

= .1016.689.

SSE = 46991 – 1016.689

= 45974.31

$$Sy.12\ldots k = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{45974.3}{10-2-1}} = 81.04$$

Table: Anova Table

Source of Variation	SS	df	MS	F
			V.R. = 508.34 / 6569.76	= 0.17
Regression	1016.689	2	508.34	
Error	45974.31	7	6569.76	
Total	**46791-**	**9**		

For α = 0.01 and 2 and 7 degrees of freedom we have 4.74- which is larger than the computed values. So, we accept H_0 and conclude that there is no linear relationship. Since-

0.082 < 4.74 P > 0.1

To test for the bi say bi:

Ho: $\beta_1 = 0$ Ha: $\beta \neq 0$

$$t = \frac{b_1 - 0}{Sb_1}$$

Sb1 = Sy. 12…….K

= 81.04 x $\sqrt{0.0308}$ = 14.222

t = 5.359

14.222= t = 0.38,

t $\alpha/2, 7$ = t $\frac{0.01}{2}$, 7

= t 0.005, 7

= 3.499

Since the computed *t* value is lesser than the table value, we accept H_0 and conclude that there is no linear relationship.

Since- 0.38 < 3.499 P > 2 (0.1)

P > 0.2.

3.12 Computing the Confidence and Prediction Intervals

The 100 (1 - α) % confidence interval is:

$$\hat{y} \pm t\ \alpha/2, n-k-1\ Sy.12 \times \sqrt{{}^{1}/_{n} + C_{11}\ X_1^{12} + C_{22}\ X_2^{12} + C_{12}\ X_1{}^{1}X_2}$$

where $X_1 = 5$ and $X_2 = 9$, the prediction interval becomes:

$\hat{y} = 36.2 + 5.39\ (5) - 1.62\ (9)$

$= 36.2 + 26.95 - 14.58$

$= 48.57$

$Sy.\ 12 = 81.04$

$C11\ X^{12}{}_1 = 0.308\ (14.1) = 0.43428$

$C_{122}\ X^{12}{}_2 = 1.56392$

$C_{12}\ X_1{}^1\ y^1{}_2 = 0.0267\ (-21.2)$

$= -0.56604$

$n = 10.$

$$\therefore \quad 48.57 \pm 3.499 \times \sqrt{1 + 0.1 + 0.434 + 1.564 - 0.566}$$

48.57 ± 451.21

$(-402.61, 499.78).$

It does not contain the population Sample; no linear relationship is found.

3.13 The Pitfalls and Limitations of Multiple Linear Regressions

In the real sense of it, the pitfalls and limitations of multiple regressions are far beyond the scope of this study; but let us just mention them.

1. The choice of an inappropriate functional form for the estimated regression equation (i.e. linear versus non-linear relationships), referred to as functional mis-specification.
2. The extent to which two or more of the independent variables are correlated with one another- referred to as the problem of *multi-colinearity*.
3. The possibility that one of the observations on the dependent variable are themselves correlated over time, referred to as a problem of auto correlation.
4. The possibility that the prediction errors may not be constant, and instead, may be correlated with the size of the independent variables, a problem referred to as *heteroscedastic*, i.e. unequal variances (opposite, *homoscedastic*).

All these aforementioned problems can be corrected when further regression analysis is carried out, which will not be treated in this study.

3.14 Correlation Analysis

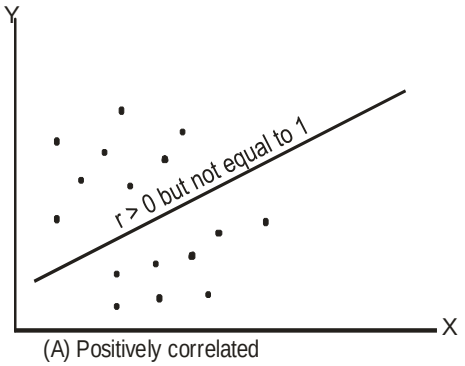
Earlier on, how to fit the best fitting line, either in simple linear or regression multiple, linear regression was mentioned to you; but we have not mentioned how weak or strong is this relationship among the dependent variable *(y)* and independent variable(s) *X*, what measure the degree or strength of closeness of the relationship between two variables is embedded in correlation analysis. When it involves two variables, we call it simple linear correlation and, when it deals with three or more variables we call it multiple correlation. We shall deliberate on both in this section.

3.14.1 Simple Linear Correlation

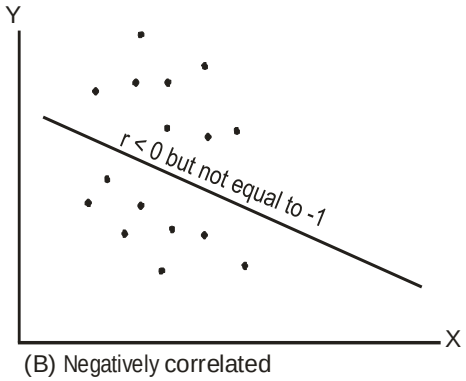
Correlation analysis provides a numerical summary measure of the degree of correlation between two variables *X* and *Y*. That is denoted by *r*, and it ranges between $0 \leq r \leq 1$.

When r = ± 1, it denotes perfectly positively correlation, as in figure (c and d)
When r = 0, it denotes no correlation, as in figure (e).

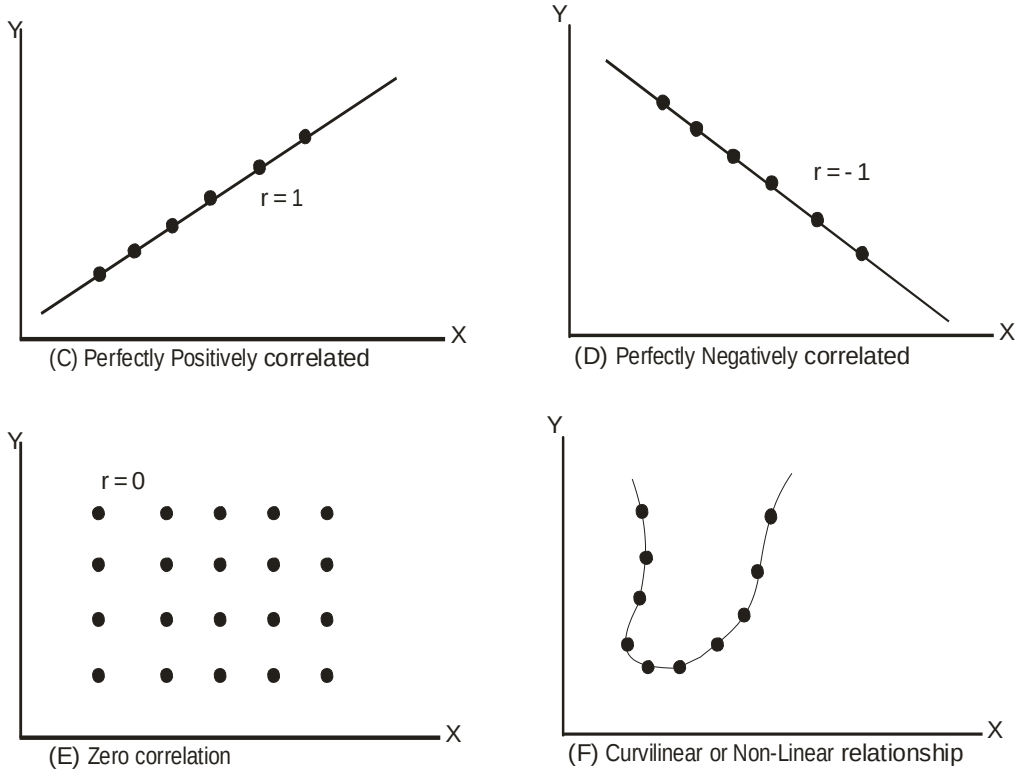
When r > 0, but not equal to 1, it denotes positive correlation (figure a)
When r < 0, it but not equal to -1, it denotes negative correlation, as in figure (b)
When r = -1, it denotes perfect negative correlation, as in figure (d)
Forms of correlation are shown in the diagrams below.



(A) Positively correlated



(B) Negatively correlated



The correlation coefficient *r*, is defined and calculated as follows:

$$r = \frac{Sxy}{SxSy}$$

$$r = \frac{\text{Cov}(X,y)}{\text{Standard deviation X multiply by standard deviating Y}}$$

$$r = \frac{\Sigma (x_1 - \bar{x})\ (y_i - \bar{x})}{\sqrt{\Sigma(x_1 - x)^2} \quad X \quad \sqrt{\Sigma(y_1 - y)}}$$

$$r = \frac{n \ \Sigma xy \ - (\Sigma x)\ (\Sigma y)}{\sqrt{n \ \Sigma x^2 \ - (\Sigma x)^2 \ (n \ \Sigma y^2 - (\Sigma y)^2}}$$

This coefficient is often referred to as the product moment correlation coefficient or Pearson's correlation coefficient.

3.16 Coefficient of Determination (R^2)

The coefficient of determination is simply the squared value of the correlation coefficient above, i.e. r^2. This simply measure the total variations in *y* (dependent variable) which is explained by the independent variable *X*. it shows how closely do the actual points cluster

around the regression line. How 'good' is the fit? Hence, it is ,deliberately, called goodness –of-fit.

$$r^2 = \frac{\text{explained variation in y}}{\text{Total variation y}} = \frac{\Sigma(y_c - \bar{Y})^2}{\Sigma(y_i - \bar{Y})}$$

$$= \frac{SSR}{SST}$$

$$= \frac{\text{Sum of square of regression}}{\text{Total sum of square}}$$

It is easily calculated thus- $r^2 = \frac{(n\Sigma xy - (\Sigma x)(\Sigma y)^2}{(n\Sigma x^2 - (\Sigma x)^2 (n\Sigma y^2 - (\Sigma y)^2}$

Example 6

It is postulated that the amount people save depends on their income. The following data show the average weekly savings *y (£)* and the average weekly income *X (£)* of people indifferent income groups.

X	19	22	27	30	36	43	47	51	61	64
Y	1.0	1.4	1.8	2.4	3.0	3.8	4.3	4.5	5.8	6.3

(a) Calculate the coefficient of correlation (b) How much of the variation in savings is explained by the variation in income?

Solving

Table 2.5

n	X	Y	Xy	X^2	Y^2
1	19	1.0	19	361	1
2	22	1.4	30.8	484	1.96
3	27	1.8	48.6	729	3.24
4	30	2.4	72.0	900	5.76
5	36	3.0	108.0	1296	9.00
6	43	3.8	163.4	1849	14.44
7	47	4.3	220.9	22.9	18.49
8	51	4.5	229.5	2601	20.25
9	61	5.8	353.8	3721	33.64
10	64	6.3	403.2	4096	39.69
10 N=10	$\Sigma x = 400$	$\Sigma y = 34.3$	$\Sigma xy = 1630.4$	$\Sigma x^2 = 18246$	$\Sigma y^2 = 147.47$

(a) $r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{(n\Sigma x^2 - (x)^2 (n\Sigma y^2 (n\Sigma y)^2}}$

$r = \frac{10 \times 1630.4 - (400)(34.3)}{}$

$$r = \frac{\sqrt{(10 \times 18246 - (400)^2)\ (10 \times 147.47 - (34.3)^2)}}{16{,}304 - 13720}$$

$$\sqrt{(182.460 - 160{,}000)\ (1474.7 - 1176.49)}$$

$$r = 0.9985$$

There is high, positive correlation.

(b) Coefficient of determination

r^2 = 99.69% . (The square of its coefficient of correlation 99.69% of the total variations in *Y* as explained by the independent variable *X*. The remaining 0.31 is due to other extraneous variable not accounted for.

3.15 Significance of the Correlation Coefficient ('R')

It becomes essential to test for the significance of the strength or weakness of the measure of relationship between *(x)* and *y* variables. We usually denote this as *P* (pronounced *rho*).

So that-

H0: $P = 0$ Ha: $P \pm 0$. or this against one sided tailed-test.

The test statistics:

$$t = \frac{r - P}{Sr}$$

Where $Sr = \sqrt{\frac{1-r^2}{n-2}}$, which will possess a *t* distribution with a – 2 degrees of freedom

$$t = \frac{r - P}{\sqrt{\frac{1-r^2}{n-2}}}$$

Example 7

Using the question, does average weekly income *(£)* have a significant influence on average weekly savings *(£)* . Let $\alpha = 0.05$.

Solution

Ho: $P = o$ Ha: $P \pm 0$

$\alpha = 0.05$ $t \alpha / 2\ n -$, $t0.025, 8 = 2.306$

$$t = \frac{0.9985 - 0}{\sqrt{\frac{1-0.9985}{10-2}}} = \frac{0.9985}{0.019358} = 51.58$$

Since *t cal* > tab we reject H_0 and affirm that average weekly income *(£)* have a significant influence on average weekly savings(£).

Since 51.58 > 2.306

$$P < 2 \quad (0.005)$$
$$P < 0.001$$

3.16 Spearman's Rank Correlation Coefficient ('r_s')

The spearman's rank correlation is a non-parametric alternative to the parametric Pearson's product moment correlation coefficient. What spearman's does is to rank two sets of samples, for instance-

$$X_1 \quad X_2 \quad X_3\ldots\ldots, X_n \text{ and}$$
$$Y_1 \quad Y_2 \quad Y_3 \ldots.., Y_n$$

In so doing, the strength and weakness of this pair ranking can be determined using the formula:

$$r_s = 1 - \frac{6\Sigma d_i^2,}{n(n^2 - 1)}$$

where we have tie of ranking, we add *1/12 (t3 – t) to* ΣD^2 , where *t* is the number tie observations, *d* is the difference between the pair samples and *n* is the number of observations.

The hypotheses to be tested are:

Ho: $P_s = 0$, Ha: $P_s \pm 0$ = (or the one-sided test $P_s >$ on $P_s < 0$)

If *n* is greater than 30, one may compute $Z = \sqrt{rs^{n-1}}$ and use the table of normal distribution to obtain critical values.

Example 8

A physiologist studying mortality of guinea pigs recorded the percentage mortality *y* against a variable *X* as follows:

X	1	2	3	4	5	6	7
Y	26	14	12	18	20	28	40

Compute: (a) the strength of the relationship using spearman's rank (b) test the significance of the relationship. Let $\alpha = 0.01$.

Table 2.6

No	**X**	**Rx**	**Y**	**Ry**	**d=Rx-Ry**	**d^2**
1	1	1	26	5	-4	16
2	2	2	14	3	-1	1
3	3	3	12	2	1	1
4	4	4	8	1	3	9
5	5	5	20	4	1	1
6	6	6	28	6	0	0
7	7	7	40	7	0	0
						Σd^2 = 28

$$r_s = 1 - \frac{6\,\Sigma d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \times 28}{7(49 - 1)} = 0.5$$

(b) H0: $P = 0$ Ha: $P \pm 0$

$\alpha = 0.01$

$t0.01/_2, 5 = t0.005, 5 = 4.032$

$$t = \frac{05 - 0}{\sqrt{\frac{1 - r_s^2}{}}} = \frac{0.5 - 0}{\sqrt{\frac{1 - 0.5^2}{}}} = \frac{0.5}{\sqrt{0.75}}$$

$$= \frac{0.5}{0.3873}$$

$$= 1.291$$

Since, *tcal < table* value- i.e. *1.291 < 4.032*, is accept H_0 and conclude that there is no relationship between the two variables.

Since 1.291 < 4.032

P > 2 (0.1) $\Longrightarrow$ P > 0 .2.

Example 9

A company attempts to evaluate the potential for a new bonus plan by selecting a random sample of ten sales persons to use the bonus plan for a trial period. The weekly sales volumes before and after implementing the plan are shown below. Are the rankings of the individual sales persons significantly different between the two schemes? Let $\alpha = 0.05$.

Table 2.7

Sales person	Weekly sales (before)	Weekly sales (after)
1	30	36
2	24	28
3	36	38
4	30	36
5	32	36
6	45	37
7	28	42
8	37	39
9	34	34
10	26	30

Ranks

Solving:

Table 2.8

No	Rx	Ry	d= Rx – Ry	d^2
1	6.5	6	0.5	0.25
2	10	10	0	0
3	3	3	0	0
4	6.5	6	0.5	0.25
5	5	6	1	1
6	1	4	-3	9
7	8	1	7	49
8	2	2	0	0
9	4	8	-4	16
10	9	9	0	0
				$\Sigma d^2 = 75.5$

Tied Ranks

$$r_s = 1 - 6 \left(\Sigma d^2 + \frac{(t3 - t)}{12} + \frac{(t3 - t)}{12} \right)$$

$$r_s = 1 = 6 \quad (75.5 + \frac{(8-2)}{12} + \frac{(7-3)}{12}$$

$$\frac{}{10 \times 99}$$

$$= 1 - 6 \left(\frac{75.5 + 0.5 + 2.0}{990}\right)$$

$$r_s = 0.527$$

Test H_0: $P = 0$ Ha: $P \pm 0$.

$$t = \frac{0.527 - 0}{\sqrt{\frac{1 - 0.527^2}{}}} = \frac{0.527}{}$$

$$\sqrt{\frac{1 - r^2}{}}$$

$$= \quad 1.754$$

Since $r_s < 3.355$, we cannot reject *Ho*. We must conclude that the rank correlation is not significantly different from zero at 5% level. Since, $1.397 < 1.754 < 3.355$, $0.1 > 0.05$.

3.17 Simple Correlation Coefficient

The simple correlation coefficient can be given as:

$$r = \frac{n\Sigma X_1 Y - (\Sigma X_1)(\Sigma Y)}{\sqrt{(n\Sigma X_1^2 - (n\Sigma X)^2 (n\Sigma Y) - (\Sigma Y)^2)}}$$

This coefficient is between the pair of variables *(y/X)*. This is often denoted as r_{11}. By extension, we define the simple correlation of three variables, where *y*, is the dependent variable *X2* and *X3* are independent variables.

Thus:

$$r_{12} \quad \frac{n\Sigma X_2 Y - (\Sigma X_2)(\Sigma Y)}{\sqrt{(n\Sigma X_2^2 - (\Sigma X_2)^2 \times (n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$r_{13} \quad \frac{n\Sigma X_3 Y - (\Sigma X_3)(\Sigma Y)}{\sqrt{(n\Sigma X_3^2 - (\Sigma X_3)^2 - (\Sigma Y^2 - (\Sigma Y)^2}}$$

Example 10

Use the data in an earlier example to compute the simple correlation coefficient.

$$r_{y1} = \frac{n\Sigma X_1 Y - (\Sigma X_1)(\Sigma Y)}{\sqrt{(n\Sigma X_1^2 - (\Sigma X)^2 \times (n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{10 \times 4673 - 67 \times 681}{\sqrt{(10 \times 463 - (67)^2 (10 \times 44735 - (681)^2}} = \frac{1103}{1233.39} = 0.8943$$

$$r_{y2} = \frac{n\Sigma X_1 Y_2 - (\Sigma X_1)(\Sigma Y_2)}{\sqrt{n\Sigma X_1^2 - (\Sigma X_1)^2 - (n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{10 \text{ x } 153 - (67) \text{ x } (26)}{\sqrt{(10 \text{ x } 463 - (67)^2 (10 \text{ x} 158 - (26)^2}}$$

$$= \frac{1530 - 1742}{\sqrt{(4630 - 44898)(1580 - 676}} = \frac{-212}{\sqrt{(141)(904)}} = \frac{-212}{357.02}$$

$$= -0.5938.$$

3.18 Partial Correlation Coefficient

This is the correlation coefficient between any two of the three variables, while allowing the other one to remain constant. In our example the partial correlation coefficient between Yx_1 keeping X_2 constant can be calculated from the formula:

$$ry1.2 = \frac{ry_1 - ry_2^2 \; r_{12}}{\sqrt{(1 - ry_2^2)(1 - r_{12}^2)}}$$

Similarly-

$$ry2.1 = \frac{ry_2 - ry_1 \; r_{12}}{(1 - r^2 y_1)(1 - r^2_{12})}$$

Is the partial correlation coefficient between y and X_2 keeping X_1, constant and

$$ry12.y = \frac{ry_{12} - r_{12} \; r\, y_2}{\sqrt{(1 - r_{12}^2)(1 - r\, y_2^2)}}$$

from the earlier data.

$$ry1.2 . = \frac{ry_1 - ry_1 \, r_{12}}{\sqrt{(1 - ry_1^2)(1 - r\, y_{12}^2)}}$$

$$= \frac{0.8943 - (0.8943)(-0.5938)}{\sqrt{(1 - 0.8943^2)(1 - (0.5938)}}$$

$$= \frac{0.8943 + 0.5311}{\sqrt{(1 - 0.8943^2)(1 - (0.5938)}} = \frac{1.4254}{0.3600}$$

$$= 3.9594$$

$$Ry2.1 = \frac{ry_2 - ry_2 \, r_{12}}{\sqrt{(1 - ry_2^2)(1 - r_{12}^2)}}$$

$$= \frac{-0.8344 - (-0.8344)(-0.5938)}{\sqrt{(1 - (-0.8344)^2 (1 - (-0.5938)^2}}$$

$$= \frac{-0.8344 - 0.4955}{}$$

$$= \frac{-1.3299}{\sqrt{(-0.3038)\ (0.6474}} = -2.9986$$
$$0.4435$$

$$r12.y = \frac{r_{12} - r_{12}\, ry_2}{(1 - r_{12}^2\,(1 - ry_2^2)} = \frac{-0.5938 - (-0.5938)\,(-08344)}{\sqrt{(0.6474)\ (0.3038)}}$$

$$= \frac{-0.5938 - 0.4955}{0.4435}$$

$$= \frac{-10893}{0.4435} = -2.4561$$

The tables of simple correlation and partial correlation can be given thus:

Correlation table

Table 2.9

	Y	**X_1**	**X_2**
Y	1	0.8943	-0.8344
X1	0.8944	1	-0.5938
X2	- 0.8344	-0.5938	1

Partial correlation table

Table 2.10

	Y	**X_1**	**X_2**
Y	1	+3.9594	-2.9986
X_1	3.9594	1	-2.4561
X_2	-2.9986	-2.4561	1

3.19 Multiple Correlation Coefficient

The coefficient of multiple determination is defined as:

$$R^2y.12 = \frac{SSR}{SST}$$

Where *SSR* is sum of squares of regression; and *SST* is total sum of squares.

$$SSR = \Sigma(\hat{Y}j - \acute{Y})^2 = bi\ \Sigma x^1y^1 + b_2\,\Sigma x_2{}^1y^1 + \Sigma X^1y_1$$

$$SST = \Sigma(Yi - \bar{Y})^2 = \Sigma y^{12}{}_i$$

$$= \Sigma y^2{}_i - \frac{(\bar{Y})}{n}$$

In the example given in this book i.e. from our Anova table.

$$R^2y.12 = \frac{1016.689}{46991.0}$$

= 0.0216 = 2.16%.

Thus, 2.16% of the total variable in *Y* is explained by the regression *Y* or X_1 and X_2; which means *R2y* is a measure of the goodness-of-fit of the regression plan to the observed points. The coefficient or multiple correlation is simply carried out by taking the square root of the coefficient of multiple determination i.e.-

$$Ry.12 = \sqrt{\frac{SSR}{}} = \sqrt{0.0216} = 0.147$$

SELF-ASSESSMENT EXERCISE

Three examiners were asked to grade ten candidates in an essay competition, and the following grades were obtained.

Candidate	A	B	C	D	E	F	G	H	I	J
Examiner I	55	60	73	45	67	51	78	48	64	57
Examiner II	62	58	66	52	71	60	75	52	69	54
ExaminerIII	60	61	57	65	61	64	60	63	65	55

By calculating the coefficient of rank correlation between all pairs of examiners, find which examiner you think disagrees most with the others.

4.0 CONCLUSION

You have seen the inter-dependence between regression and correlation analysis. Why it is necessary to know the relationship between two or more variables, it is equally sufficient to explain the strength of their relationship. This is embedded in the study of regression and correlation analysis. Your understanding of this unit is pertinent to your various fields of study. In regression, your main concern is to estimate the regression coefficients and you can see from the unit that it is a straight forward business. Once this is achieved, you can now use it to predict by interpolation. Testing the strength of regression relationship is also a simple thing, as you can see from the unit.

5.0 SUMMARY

In this unit, you have learnt how to determine the relationship between two or more variables having one dependent variable. You can now plot a scatter diagram of this relationship, and once this is plotted you can determine the relationship through the eye's fit method. You can also use the least squares method to compute the constant and regression coefficient of the relation $Y = \alpha + \beta x$. You can at the same routine evaluate the strength of the relationship through correlation analysis.

6.0 TUTOR MARKED ASSIGNMENT

i. In an attempt to determine to what extent investment expenditure *I* is influenced by the rate of interest R, the data given below were collected over an eight year period.

Year	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Investment (t (Σm)	2.1	1.8	1.8	2.2	2.8	4.1	3.6	3.1
Average rate of interest (R) %	9.6	9.9	90.5	9.8	12.1	7.7	9.5	9.2

(a) Is the expected sign of the regression slope coefficient positive or negative?
(b) By how much does investment change in response to *I* percentage point decline in the rate of interest.
(c) Does the rate of interest have a significant influence on the level of investment?
(d) How much of the variance in investment is explained if variable in the rate of interest?

ii. Find the multiple – regression equation for the sample data below.

Yield Y	**Fertilizer application**	**Index of soil quantity**
50	38	50
52	39	50
56	39	54
59	41	56
62	44	56
64	42	60
68	43	64
69	46	63
70	48	62
71	47	60

iii. In the table below: (a) compute the multiple correlation coefficient (b) Test H0 that Py.12 = 0 (c) compute the simple and partial correlation coefficients. (d) Test H0 that Py1.2 = 0 (Let α =0.05) for all lest and determine the P-values).

Plot	Age (X_1)	Fertility (X_2)	Yield (Y)
1	15	60	0.8
2	20	76	1.2
3	24	84	2.3
4	38	85	3.2
5	44	86	4.3
6	48	85	4.8
7	56	86	5.0
8	62	86	5.5
9	63	87	6.3
10	64	87	7.2
Total	434	822	40.6

Σx^2_1, = 21, 930 Σy^2_i = 205.92 $\Sigma y_{2i}\ y_i$, = 3459.7
Σy^2_2 = 68,208 $\Sigma y_{1;} yi$ = 2111.1 $\Sigma y_i X_2$, = 36,727

7.0 REFERENCES/FURTHER READING

Brian, P.M. & Philip, M.N. (2006). *Applied Statistics for Public Policy*. New Delhi: Prentice-Hall.

Daniel, T. (1990) *Business Statistics*. (2nd ed.). Haughton Milton Company International.

James, T. & Benon, P.G. (1988). *Statistics for Business and Economics*.(4th ed.). United States: Dellen Publishing Company Inc.

UNIT 3 ANALYSIS OF VARIANCE

CONTENTS

1.0 INTRODUCTION

You have already learnt how to test the hypothesis of the mean of two populations, using the *Z* and *T* test. However, we have not done so to the mean of three or more population. Therefore, this unit (and the next) will educate you on how to compare the mean of three or more population. This is referred to as the analysis of variance.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- describe analysis of variance
- evaluate one way analysis of variance
- test for significant difference among means.

3.0 MAIN CONTENT

3.1 Analysis of Variance (*ANOVA*)

Analysis of variance is most appropriate when comparing the mean of two or more populations. This is often called *ANOVA.* What we do here is that we partition the variance into two components, namely:

(a) variance due to treatment effect
(b) variance due to random error

Then TSS = TR SS + ESS

We may display experimental survey data that are to be analysed by one-way *ANOVA* thus-

X_{ij} = the *ith* observation that receives the ith treatment, l = 1, 2 …., nj = j = 1, 2, …. K.

Xj = $\underline{T.j}$ = Means of the jth column
nj

T… =

$$T... = \sum_{j=i}^{K} T.\ j = \sum_{i=l}^{n} \sum_{i=l}^{K} X_{ij} = \text{total or}$$

That is also referred to as the grand mean.

$$\text{Let } n = \sum_{J=l}^{K} nj$$

Table 3.1: *Sample Data for Analysis by ANOVA Population*

	1	2	3	.	.	.	K	
	X_{11}	X_{12}	X_{13}	.	.	.	X_{1K}	
	X_{21}	X_{22}	X_{23}	.	.	.	X_{2K}	
	X_{31}	X_{32}	X_{33}	.	.	.	X_{3K}	
	.	.	.	..	.	.	.	
	.	.	.	.	.	.	.	
	.	.	..	.	.	.	.	
	Xn_1^1	Xn_2^2	Xn_3^3	.	.	.	Xn_K^K	
Total	T.1	T.2	T.3	.	.	.	T.K	T..
Mean	X.1	X.2	X.3	.	.	.	X.K	X..

The total sum of square (TSS)- this is the observation of squares or the deviation of each observation from the means of all the observations taken together. We define this total sum of sequence as:

$$TSS = \sum_{i=l}^{n} \sum_{i=l}^{K} (X_{ij} - \overline{X}...)^2$$

We may re-write this equation thus:

$$TSS = \sum_{i=l}^{n} \sum_{i=l}^{K} X^2_{ij} - \left(\frac{\sum_{i=l}^{n} X^2}{ij} \right)$$

$$n = K_{nj}$$

If we let

$$\frac{\left(\sum_{j=i}^{K}\sum_{l=1}^{nj} Xij\right)^2}{n = knj} = \frac{T_{oo}^2}{N}, \text{ where } N = knj$$

known as the correction term, C.

Then

$$TSS = \sum_{}^{K}\sum_{}^{nj} X_{ij}^2 - \frac{T_{oo}^2}{N}$$

$$TSS = \sum_{j=i}^{K} nj\left(\overline{X}_j - \overline{X}_{oo}\right) = T_{.1}^2 + T_{.2}^2 + \ldots + \frac{T_{nk}^2}{K} - \left(\sum_{n=kjn}^{K}\sum^{nj} Xij\right)^2$$

Which, when all grouped are equal in sizes reduced to:

$$TrSS = \sum_{j=i} T_{.j}^2 - \frac{T_{oo}^2}{N}$$

$$ESS = \sum_{j=i}^{K}\sum_{i=l}^{nj}\left(X_{ij} - \overline{X}_{.j}\right)^2$$

We may not need to border much about the right hand side, since

TSS = TRSS + ESS and
ESS = TSS – TRSS

Table 3.2: ANOVA Table

SOURCE OF VARIATION	SUM OF SQUARE (SS)	Df	MEAN SQUARE (MS)	VR
Treatment	Trss	k-1	Trss/k= Mtrss	F = Mtrss / Mse
Error	ESS	N-K	SSE/N-K = MSE	
TOTAL	TSS	N – 1		

The test statistic as shown in table has
F = (TRSS/KY) = MTRSS, with K-1 and N – K Degree of freedom.
ESS/N-K MSE

Example 1

In testing customer acceptance of a new product, 4 different computer displays are used. Thirty-six stores, matches on all relevant criteria, are selected, with each display being used in 9 of the stores. Total sales (coded) at the end of week are as follows. At the 0.05 level of significance, test the null hypothesis of no difference among the four means.

A	**B**	**C**	**D**
5	2	2	6
6	2	2	6
7	2	3	7
7	3	3	8
8	3	2	8
6	2	2	8
7	3	2	6
7	3	3	6
6	2	3	6

Solution

1. Model
   $Yij = \mu + \tau_j + \acute{\varepsilon}_{ij}$
2. Assumption: fixed effect, four different counters are used, 9 stores for each counter.
3. Hypothesis:
   H_o: $\mu A = \mu B = \mu C = \mu D$
   Ha: At least, one equality does not hold.
4. Calculation-

Table 3.3

A	B	C	D
5	2	2	6
6	2	2	6
7	2	3	7
7	3	3	8
8	3	2	8
6	2	2	8
7	3	2	6
7	3	3	6
6	2	3	6

Total T.j: 59 22 21 61 Too = 164

T^2_j = 3481 484 484 3721 $=\Sigma T^2_j = 8170$

$\sum_{j=i}^{K} \sum_{i=l}^{nj}$ 393 56 56 421

$\sum_{j=i}^{K} \sum_{i=l}^{nj} = 926$

$$TSS = \sum_{}^{K} \sum_{}^{nj} X^2_{ij} - \frac{T^2_{oo}}{nk}$$

$$= 926 - \frac{(164)^2}{36}$$

$$= 926 - 747.11$$

$$= 178.9$$

$$TrSS = \sum_{j=i} T^2_{.j} - \frac{T^2_{oo}}{N}$$

$$= \frac{8170}{9} - 747.11$$

$$= 907.8 - 747.11$$

$$= 160.7$$

SSE = TSS – Trss
= 178.9 – 10.7
= 18.2

Anova Table Fractions

SOURCE	DF	SS	MS	
Treatment	3	160.7	53.57	F = 53.57 = 94.15
Error	32	18.2	0.569	
Total	35	178.9		

$F_{critical}$, F3, 32, 0.05 = 2.92

Conclusion- Since 94.15> 2.92 we reject Ho: Since 94.15>63.32, P < 0.05, and

Conclude that the means are different
The problem now is which of the means are really different. The only measure to use in deciding this is by making pair wise companions. Several of these computations have been alluded to in statistical textbooks, such as bonferroni, Turkey's HSD test, Scheffy etc.
Let us demonstrate Turkey's *HSD* test whose statistics is given as:

$$HSD = q_{\alpha, k, n-k,} \sqrt{\frac{MSE}{nj}} \quad \text{for}$$

equal sizes, where *x* = level of significance, *K* is the number of means in the experiment, *n* is the number of observations in a treatment, *MSE* is error within means square from the *ANOVA* table and *q* is the percentage point of the students range with α, *K and N – K* degree of freedom
However, for an unequal size-

$$HSD = q_{\alpha, k, n-k,} \sqrt{\frac{MSE}{nj}} \quad \text{for}$$

where, *nj* is the smallest of the two sample sizes associated with the two sample means that are to be compared.

For our example above, the means are-
6.56 2.44 2.44 6.78
Arrange them in a descending order of magnitude:

	$\overline{X}$	$\overline{X}$	$\overline{X}$	$\overline{X}$
$\overline{X}_D = 6.78$	—	0.22	4.34	4.34
$\overline{X}_D = 6.56$	—	—	—	0
$\overline{X}_D = 2.44$	—	—	—	—
$\overline{X}_D = 2.44$	—	—	—	—

Find their absolute value difference

$$HSD = q_{0.05, 4, 32} \sqrt{\frac{0.569}{9}}$$

$$= 3.84 \times \sqrt{\frac{0.569}{9}}$$

$$= 0.966$$

We compare the differences between means shown above with 0.9660.

$$\left| \overline{X}_D - \overline{X}_C \right| = 0.22$$

$$\left| \overline{X}_A - \overline{X}_B \right| = 0$$

Pair means which exceeds 0.966 is declared significant and is rejected, except 0.22 and 0.

3.2 Randomised Complete Block Design

This is an *ANOVA* design which the units (called experimental units) to which the treatments are applied are sub-divided into homogeneous groups called blocks, so that the number of experimental unit in a block is equal to the number (or some multiple numbers) of treatments being scheduled. The treatments are then assigned at random to the experimental units within each block and each treatment appears in every block and each block receives every treatment.

The calculation is such that we determined the sum of squares of block along the row and the sum of sequence of square of treatment along the columns. Hence, it is called too ways analysis of variance.

The model is given as:

$Y_{ij} = \mu + \beta_i + + Tj + \Sigma ij$
$l = 1, 2, \ldots, nj = l, 2, \ldots, K$

Table of sample of values for the randomised complete block design
Treatments

Table 3.4

Blocks	1	2	3	4	…	K	Total	Mean
1	X_{iii}	X_{12}	X_{13}	X_{14}	…	X_{1k}	T_1	X_1
2	X_{21}	X_{22}	X_{23}	X_{24}	…	X_{2k}	T_2	X_2
3	X_{31}	X_{32}	X_{33}	X_{34}	…	X_{3k}	T_3	X_3
.	.	.	.	.	.	.	.	.
.	..	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
N	Xn1	Xn2	Xn3	Xn4	Xn4	Xnk	T.	Xn
Total	T.1	T.2	T.3	T.4	…T.K	Tn	T..	
Mean	X.1	X.2	X.3	X.4	…NT	X…		

From the above
TSS = BSS + TRSS + ESS
ESS = TSS – BSS – TRSS

$$TSS = \sum^{K} \sum^{nj} X^2_{ij} - \frac{T^2_{oo}}{nk}$$

$$BSS = \sum_{j=i}^{K} \frac{N Ti^2}{} - \frac{T^2_{..}}{nk}$$

$$TrSS = \sum_{j=i} \frac{T^2_{.j}}{n} - \frac{T^2_{oo}}{N}$$

Table 3.5: ANOVA Table for the Randomised Complete Block Design

Source	SS	Df	MS	VR
Treatments	Trss	(K – 1)	MTRSS = TRSS/ (K-1)	MTRSS / MESS
Blocks	BSS	(n-1)	MBSS = BSS/(n-1)	
Errors	ESS	(n-1) (k-1)	MESS=ESS /(k-1)(n-1)	
Total	TSS	Kn-1		

Example 2

The nursing supervisor in a local health department wished to study the influence of time of days on length of home visit by the nursing staff. It was thought that individual differences among nurses might be large, so the nurse was used as a blocking factor. The nursing supervisor collected the following data.

Table 3.6: Length of Home Visit by Time of Day

Nurse	Early morning	Late morning	Early afternoon	Later afternoon
A	27	28	30	23
B	31	28	27	20
C	35	30	24	30
D	20	18	20	14

Do these data provide sufficient evidence to indicate differences in length of home visit all through the different times of the day?

let $\alpha = 0.05$
Ho: $U_A = U_B = Uc = U_D$
H_a: At least, one equality does not hold.

Table 3.7

NURSE	T_1	T_2	T_3	T_4	TOTAL
A	27	28	30	23	108
B	31	30	27	20	108
C	35	38	34	30	137
D	20	18	20	14	72
Total t.j	113	114	111	87	425 = T.0
$T^2.j$	12769	12996	123.21	7569	
$\Sigma\Sigma n^2_{ij}$	33155	3452	3485	2025	

$$TSS = \sum_{j=i} T.^2 j - \frac{T^2 o\, o}{N}$$

$$= 3315 + 3452 + 3185 + 2025 - \frac{(425)^2}{4x4}$$

$$= 1197 - 11289.063$$

$$= 687.937$$

$$TrSS = \sum_{j=i} \frac{T.^2 j}{n} - \frac{T^2..}{}$$

$$= 113 + 114 \ldots + 87 - \frac{(425)^2}{}$$

$$= \frac{12769 + 12321 + 7569 - 1129.063}{4}$$

$$= 11413.75 - 11289.063 = 124.687$$

$$BSS = \frac{\sum_{j=i}^{N}}{K} - \frac{T^2 o\, o}{nk}$$

$$= \frac{108^2 + 108^2 + \ldots 72^2}{4} - 11268.063$$

$$= 11820.25 - 11289.063$$

$$= 531.187$$

$$ESS = 687.937 - 531.187 - 124.687$$
$$= 32.068$$

Table 3.8: ANOVA Table

SOURCE	SS	Df	MS	F
Treatment	124.687	3	41.56	VR=41.56 3.563
Blocks	531.187	3	177.06	
Error	32.063	9	3.563	
Total	687.937			

F tab (3, 9), 0.05 = 3.86
Since, $F_{cal} = F_{critical}$, it is significant; hence, we reject H_0, that it provides sufficient evidence to indicate the difference in length of home visiting through the different times of the day. Since,11.66>8.72, P < 0.005.

3.3 Latin Square Design

In Latin square design, we isolate two sources of extraneous variations from error variance. We assign one source of the extraneous variation to the columns of the square, and another source to the rows of square; and we assign the treatment, which are designated by Latin letters, in such a way that each treatment occurs, once (and only once) in each row and each column.

The number of rows, the number of columns and the numbers of treatments, therefore, are all equal- i.e. 4 x 4, 5 x 5, etc. Since we have columns, rows and treatments sum of square, it is often called three-way-analysis of variance.

The model is given as:

$Y_{ijk} = \mu + \alpha_i + \beta_i + \lambda_k + \Sigma_{ijk}$
The design for the Latin square is given in the table below.

Table 3.9: Three-Way-Analysis of Variance Column

ROWS	1	2	3	4	TOTALS	ROW MEANS
1	X_{11t}	X_{12t}	X_{13t}	$X_{14t....r}$	$T_{1...}$	$\bar{X}_{1...}$
2	X_{12t}	X_{13t}	X_{23t}	$X_{24t....x_{2r}t}$	$T_{2.....}$	$\bar{X}_{2....}$
3	X_{31t}	X_{32t}	X_{33t}	$X_{34t....x_{3r}t}$	$T_{3.....}$	$\bar{X}_{3....}$
4	X_{41t}	X_{42t}	X_{43t}	$X_{44t....x_{4r}t}$	$T_{4.....}$	$\bar{X}_{4....}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
r	X_{r1t}	X_{r2t}	X_{r3t}	$X_{r4t....x_{1r}t}$	$T_{r.....}$	$\underline{W}$ $X_{r....}$

Colum Total	T_1	$T_{.2.}$	$T_{.3.}$	T_4	T_r	$T_{..}$
Colum Mean	$\overline{X}_{.1,}$	$\overline{X}_{.2.}$	$\overline{X}_{.3.}$	$\overline{X}_{.4.}$	$\overline{X}_r$	

From the above table:

TSS = RSS + CSS + TRSS + ESS
ESS = TSS – RSS - CSS - TRSS

$$TSS = \sum_{j=i,\,k=1}^{K} \sum^{nj} X^2_{ij} - \frac{T^2_{ooo}}{r^2}$$

$$RSS = \sum_{j=i}^{r} \frac{T^2_{oo\,o}}{r^2} - \frac{T^2_{ooo}}{r^2}$$

$$CSS = \sum_{j=i}^{r} \frac{T_j^2}{r} - \frac{T^2_{ooo}}{r^2}$$

$$TrSS = \sum_{K=i}^{r} T_{.}^2 K - \frac{T^2_{..}}{N^2}$$

Table 3.10: The Anova Table

SOURCE	SS	DF	MS	F
Row	RSS	(r – 1)	Msr=RSS/(R-I)	MTRSS / MESS
Column	CSS	(r – 1)	MSC = CSS/(r-1)	
Treatments	Trss	(r-1)	MTRSS = TRSS/(R-1)	
Error	ESS	$\frac{R^2 - 3r + 2}{R^2 - 1}$	MESS = ESS/r^2 - 3r +2	
Total				

Example 3

In a study designed to evaluate the performance of students after eliminating methods and teacher effects and the treatments gain in the students pre-test and post-test scores. The table below shows the result of the experiment.

Table 3.11

TEACHER	T1	T2	T3	T4
1	A(12)	B(10)	C(10)	D(10)
2	B(10)	A(12)	D(12)	C(10)
3	C(10)	D(11)	B(10)	A(12)
4	D(11)	C(10)	A(13)	C(10)

Test, at the 0.05 level of significance, the null hypothesis of no difference between treatment means.

Solution

H_0: $\mu_A = \mu_B = \mu_C = \mu_D$

Table 3.11: Row Column (Method)

TEACHERS	T1	T2	T3	T4	ROW TOTAL	ROW MEANS
1	A=12	B=10	C=10	D=12	44	11
2	B=10	A=12	D=12	C=10	44	11
3	C=10	D=12	B=10	A=12	43	10.75
4	D=11	C=10	A=13	B=10	44	11
TOTAL	43	43	45	44	175	

COLUMN MEANS	10.75	10.75	11.25	11	10.94
T^2J	1849	1849	2025	1936	=7659=T…
ΣX^2jk	465	465	513	488	$T_j^{\ 2}$
					1936
					1936
					1849
					1936

Table 3.12: Treatment Total

	A	B	C	D
	12	10	10	12
	12	10	10	12
	12	10	10	11
	13	10	10	11
T..K	49	40	40	46
T^2..K	2401	1600	1600	2116 = ΣT^2..k = 7717

MEAN (X …k) 12.25, 4, 4, 11.5

$TSS = 12^2 + 10^2 + \ldots 10^2 - \frac{(175)^2}{16}$

$= 465 + 465 + 513 + 488 - 1914.06$

$= 1931 - 1914.06$

$= 16.94$

$RSS = \frac{44^2 + 44^2}{4} - 1914.06$

$= 1936 + 1936 + 1849 + 1936 - 0.1914.06$

$= 1914.25 - 1914.06 = 0.19$

$CSS = 43^2 + 43^2 + \ldots + 44^2 - 1914.06$

$= 1849 + 1849 + 2005 + 1936 - 1914.06$

$= 1914.74 - 1914.06$

$= 0.69$

$TRSS = 49^2 + \ldots + 46^2 - 1914.06$

$= 15.19$

$ESS = TSS - RSS - CSS - TRSS$

$= 16.96 - 0.19 - 0.69 - 15.19$

$= 0.89$

Table 3.13: ANOVA Table

	SS	DF	MS	CR
ROW (TEACHER)	0.19	3	0.0633	$FC=\frac{5.063}{0.1483} = 34.14$
COLUMN (METHODS)	0.69	3	0.23	
TREATMENTS	15.19	3	5.063	
ERROR	0.89	6	0.1483	
TOTAL	16.94	15		

$F_{critical} = F_3\ 6, 0.05 = 4.76.$

Since, $F_{critical} < F_{cal}$, it is highly significant. We shall reject H_o and conclude that there are differences among treatment means. Since, 34.14 > 12.92, P<0.005.

SELF-ASSESSMENT EXERCISE

In a study designed to evaluate 4 brands of lubricating oil, a Latin square design is used where the columns represent the four seasons of the year and, the rows represent 4 makes of vehicle. The variable of interest is the fuel consumption in gallon per miles travelled. The results are as

follows; test at the 0.05 level of significance the null hypothesis, no difference between treatment means.

SEASON

Vehicle	F	W	Sp	Su
1	A(12)	B(10)	C(10)	D(12)
2	B(10)	A(12)	D(12)	C(10)
3	C(10)	D(11)	B(10)	A(12)
4	D(11)	C(10)	A(13)	B(10)

4.0 CONCLUSION

You can now apply this unit whenever you are testing hypothesis and comparing the means of three or more population means. First and foremost, you need to understand that the basic component of analysis of variance is the treatment sum of squares and the error sum of squares. Once you are able to identify these components, you can now use the *F-test* distribution (critical point) to compare the F- statistics, which is the ratio of mean sum of squares of treatment to mean error sum of squares.

5.0 SUMMARY

You have learnt in this unit, the analysis of variance model for one way analysis of variance (ANOVA), two way analysis of variance and three ways analysis of variance - popularly called Latin square design. You can now apply this in hypothesis of three or more population means.

6.0 TUTOR-MARKED ASSIGNMENT

i. Three machines are used to perform a packaging process, but for various reasons, the hourly output of each machine varies. It is desired to test whether or not the machines are different. Six observations of hourly output for each machine have been collected, at random, and are shown below. Is there a significant difference between the three machines?

Machine 1:	31	23	29	30	34	21
Machine 2:	30	31	28	20	21	26
Machine 3:	28	34	38	35	29	37

ii. The managing director of a chain of ladies fashion boutiques wishes to test whether or not sales are affected by playing music in the shops and the type of music played. It is also thought that sales may depend on the towns in which the shops are located. An experiment is conducted in which shops play either classical

or pop music or no music at all and daily sales (₦000) recorded. A random sample of 21 shops gave the result for average daily sales as shown below.

Town	Non Music	Classical Music	POP Music
A	15	29	41
B	27	51	12
C	8	11	123
D	65	62	98
E	75	23	55
F	18	112	21
G	87	84	61

Test all the 0.1 level of significance; the null hypothesis of no difference between treatment means.

7.0 REFERENCES/FURTHER READING

Brian, P.M. & Philip, M.N. (2006). *Applied Statistics for Public Policy*. New Delhi: Prentice-Hall .

James, T. & Benon, P.G. (1988). *Statistics for Business and Economics*. (4th ed.). United States: Dellen Publishing Company Inc.

Kasumu, R. B. (2002). *Introduction to Probability Theory*. *A First Course*. JAB Publishers.

UNIT 4 ANALYSIS OF COVARIANCE (ANCOVA)

CONTENTS

1.0 INTRODUCTION

This analysis is based on inclusion of supplementary variable (covariates) into the model of analysis of variance. This lets us account for the group variation associated, not with "treatment" itself, but the covariate(s); and since we suspect a positive correlation between *x* and *y* the model turns to a regression model.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- describe analysis of covariance
- compute the analysis of covariance
- evaluate the mean of each adjusted independent variable.

3.0 MAIN CONTENT

3.1 Analysis of Covariance

Here, we shall introduce a covariate (X) into our model and turn it into a regression model, such that:
If $Y_{ij} = \mu + T_j + \Sigma_{ij}$ ---- *

$Y_{ij} = \mu + \alpha (x_{ij} - \overline{X}) + \Sigma_{ij}$ -------**
When α is the true linear regression coefficient (or slope) between *y* and *x* over all the data, *X* is the means of *X* values.
Therefore, for analysis of co-variance (ANCOVA). We combine x and y.

$Y_{ij} = \mu + T_j + \alpha (x_{ij}` - \overline{X}) + \Sigma_{ij}$

3.2 Computations

1. We compute Tssy = Total sum of squares of λ

$$= \sum_{i=1}^{n} \sum_{i=1}^{K} Y_{ij}^2 - \frac{\left(\sum_{i=1}^{n} \sum_{i=1}^{K} Y_{ij}\right)^2}{n_T}$$

2. The sum of square treatments, which determines the variability between populations of factors, n_k represents the number of factors.

$$TRSSY = \sum_{i=1}^{n} \left(\frac{\sum_{i=1}^{n} \sum_{i=1}^{K} Y_{ij}^2}{n_K} \right) - \frac{\left(\sum_{i=1}^{n} \sum_{i=1}^{K} Y_{ij}\right)^2}{n_T}$$

3. The sum of square for error determines the variability with each population or factor, n_k represents the number of samples with a given population.

$$ESSY = \sum_{i=1}^{n} \sum_{i=1}^{K} Y_{ij}^2 - \sum_{i=1}^{n} \left(\frac{\sum_{i=1}^{n} Y_{ij}^2}{n_K} \right)$$

Although, we may not need to use the formula above since
TSSY=TRSSY+ESSY
ESSY = TSSY – TRSS

4. We compute the sum of square for the covariant(s).

$$TSS_X = \sum_{i=1}^{n} \sum_{i=1}^{K} X_{ij}^2 - \frac{\left(\sum_{i=1}^{n} \sum_{i=1}^{K} X_{ij}\right)^2}{n_T}$$

$$TRSS_{Xy} = \sum_{i=1}^{k} \left(\frac{\sum_{i=1}^{n} Y_{ij}^2}{n_K} \right) - \frac{\left(\sum_{i=1}^{n} \sum_{i=1}^{K} Y_{ij} \right) \left(\sum_{i=1}^{k} \sum_{i=1}^{n} \right)}{n_T} \quad \frac{T}{n}$$

$$ESS_{XY} = TSS\ C_{XY} - TRSSC_{XY}$$

5. Adjust TSS_Y - this is achieved by finding the correlation between x and y as follows:

$$r_t^2 = \frac{(TSS\ C_{XY})^2}{TSSx\ X\ TSS_Y}$$

$$r_n^2 = \frac{(ESS\ C_{XY})^2}{ESSx\ X\ ESS_Y}$$

TSS_Y adj = TSSy $(1\text{-}r^2{}_T)$
ESS_yadj = ESS_y $(1\text{-}r^2)$
ESS_Y adj = TSSadj - ESSyadj

The mean of each population is adjusted thus:

$$M_{yadj} = M_{yi} = ESS\ C_{XY} \frac{(M_{xi} - M_{XT})}{ESSC_x}$$

The ratio becomes:

$$Fdf, dF_E = \frac{MTRSS}{MESS} = \frac{TRSS_{yad}/df_{TR}}{ESS_{yadj}/d_{FE}}$$

Example 1

The data below are for five individuals who have been subjected to four conditions of treatments represented by the groups represented by the group or lots. 1- 4λ represents some measure of an individual supposedly affected by the variations in treatments of the four lots. X represents another measure of an individual which may affect the values of y even in the four groups treatments. Determine whether or not that the y means of four differ significantly from each other after the effect of x variable have been removed.

Table 4.1: Group or Lots

1		2		3		4	
X	Y	X	Y	X	Y	X	Y
29	22	15	30	16	12	5	23

20	22	9	32	31	8	25	25
14	20	9	32	31	8	25	25
21	24	6	25	35	25	10	26
6	12	19	37	12	7	24	23

Solution

Model $Y_{ij} = \mu + T_j + \alpha (X_{ij} - \overline{X}) + \Sigma_{ij}$

Hypotheses:

H_0: $\mu_A = \mu_B = \mu_c = \mu_D$
H_a: At one of the equality does not hold.

Table 4.2

	1		2		3		4		Total	
	X	Y	X	Y	X	Y	X	Y	X	Y
	29	22	15	30	16	12	5	23		
	20	22	9	32	31	8	25	25		
	14	20	9	32	31	8	25	25		
	21	24	6	25	35	25	10	26		
	6	12	19	37	12	7	24	23	340	440
TOTAL	90	100	80	150	120	65	80	125		

$$TSSY = 22^2 + 22^2 + 20^2 \ldots + 23^2 - \frac{(440)^2}{20}$$

$$= 1{,}196$$

$$TSS_X = 29^2 + 20^2 + \ldots + 24^2 - \frac{(340)^2}{20}$$

$$= 7462 - 5780$$

$$= 1682$$

$$TSSC_{XY} = (29)(22) + \ldots (24)(23) - \frac{(340)(440)}{20}$$

$$= -192$$

$$TRSS_X = \frac{90^2 + 50^2 + \ldots 80^2}{5} \quad \frac{(340)^2}{20}$$

$$= 500$$

$$TRSS = \frac{100^2}{5} + \frac{150^2}{} + \ldots + 125^2 - \frac{(440)^2}{20}$$

$$= 790$$

$$TRSSC_{XY} \frac{(90)(100) + (50)(150)}{5} + \ldots \frac{(80)(125) - (340)\text{-}(440)^2}{20}$$

$$= -620$$

$$TSSY = TSSY - TRSSY$$

= 1,196 – 790
= 406
$ESS_X = TSS_X - TRSS_X$
= 1682 – 500
= 1182
$ESSC_{XY} = TSSC_X - TRSSC_X$
= -192 + 620
= 428

Adjust *TSSY* between correlation *x* and *y* is

$$r^2_T = \frac{(TSSC_{XY})^2}{TSS_X\ TSS_Y}$$

$$= \frac{(-192)^2}{1682 \text{ X } 1196} = 0.0183$$

$$r^2_n = \frac{(ESSC_{XY})^2}{ESS \text{ X } ESS_T} = \frac{(428)^2}{1182 \text{ X } 406} = 0.3817$$

$TSS_{Yajd} = TSS_Y (1-r_T^2)$
= 1,196 (1- 0.0183)
= 1174.08

$ESS_{Yadj} = ESS_Y (1- r^2n)$
= 406 (1-0.3187)
= 215.02
$TRSS_{adj} = TSSY_{ADJ} - ESS_{adj}$
= 1174.08 – 251.02
= 923.06

Ancova Table

Source	df	Σx^2	Σ_{xy}	Σ_y^2	Σ_y^2	df	ms
Between lots	3	500	-620	790			
Errors	16	1182	428	406	251.0215		16.73
Total	19	168	-198	1196	1174.08		18
Between treatment (adj)		923.063	307.69				

$$F\ df_T, df_F \quad \frac{MTrss(adj)}{MESS\ (adj)} = \frac{923.0613}{251.02/_{15}} = \frac{307.69}{16.73}$$

= 18.39

$F\ df_T, 0.05 = F_3, 15, 0.05$

This is highly significant at the 0.05 significant level. This means that after removing the efforts of the *x* variables, the *y* means of the four groups differ, significantly, from each other.

The mean of each population is adjusted in the following manner.

$$M_y\ (adj) = M_{yii} = \frac{ESSC_{XY}}{ESSC_X}\ (M_{yi} - M_{xi})$$

Ffor plot 1:

$$Y_i = \frac{100}{5} - 0.362\left(\frac{90}{5} - \frac{340}{20}\right)$$

Plot 2 $Y_2 = 32.52$
Plot 3 $Y_3 = 10.47$
Plot 4 $Y_4 = 25.36$

4.0 CONCLUSION

You have learnt, in this unit, the importance of analysis of covariance as a new approach to analysis of variance where regression model is added to the analysis of variance. You can now apply it in your research work.

5.0 SUMMARY

You have learnt in this unit the covariate which is included in analysis of variance which changes our relationship to regression model. Your ability to master the computations is contingent upon determining the total sum of squares that are accounted for by the variables *X* and *Y,* and the total sum of covariate between X and *Y*. Having done this, you can now adjust the treatment and error sum of squares, so as to determine *F* distribution statistics and adjust the means of the dependent accordingly.

6.0 TUTOR-MARKED ASSIGNMENT

i. An animal scientist is interested in determining the effect of four different feed plans on hogs. Twenty four hogs of a breed were chosen and randomly assigned to one of the four feeding plans for certain period. Initial weight (x) of the hogs and gains in weight (y) in pounds at the end of the experiment are given below.

Feed Plan

1		2		3		4	
X	Y	X	Y	X	Y	X	Y
30	165	24	188	34	56	41	201
27	170	31	169	32	189	32	173
20	130	20	171	35	138	30	200
21	156	26	161	35	190	35	193
33	167	20	180	30	160	28	142
29	151	15	170	29	172	36	189

Determine whether or not the mean of each of the four groups differs significantly from each other, after the effects of x variable have been removed.

7.0 REFERENCES/FURTHER READING

Brian, P.M. & Philip, M.N. (2006). *Applied Statistics for Public Policy*. New Delhi: Prentice-Hall.

James, T. & Benon, P.G. (1988). *Statistics for Business and Economics*. (4th ed.). United States: Dellen Publishing Company Inc.

Kasumu, R. B. (2002). *Introduction to Probability Theory*. *A First Course*. JAB Publishers.

MODULE 4 OPERATIONS RESEARCH

UNIT 1 INTRODUCTION TO OPERATIONS RESEARCH

CONTENTS

1.0 INTRODUCTION

In this unit, you will discover a new scientific method to problem solving. You will be able to see how you can utilise and integrate complex functional relationship into mathematical model.

2.0 OBJECTIVES

At the end of the unit, you should be able to:

- define operations research
- describe the application of operations research
- describe the phases of operation research.

3.0 MAIN CONTENT

3.1 What is Operations Research?

Operations research is the application of scientific methods to problems arising from operations involving integrated systems of men, machines and materials. It utilises the inter-disciplinary team in order to represent complex functional relationships as mathematical models for the purpose of providing a quantitative analysis.

Operations research is concerned with determining the maximum of (profit, performance or yield) or minimum of (of cost, loss or risk) of some real world objective. Operations research encompasses a wide range of problem solving techniques and methods applied in the pursuit of improved decision making and efficiency such as simulation, mathematical optimisation, queuing theory etc.

3.2 History of Operations Research

The main origin of operations research was during World War II. The military command of UK and US engaged several inter-disciplinary teams of scientists to undertake scientific research into strategic and tactical military operations. Since they were having very limited military resources, it was necessary to decide upon the most effective utilisation of available resources e.g. efficient transportation etc.

The mission of these scientists was to formulate specific plans for aiding the military commands to arrive at a decision, effectively. They were not physically involved in the fighting, but were advisors and were, significantly, instrumental to winning the war. After the war, the techniques began to be applied more widely to problems in businesses, industries and society. Application of operations research in management cut across the following areas.

1. Finance – budgeting and investments
2. Purchasing , procurement and exploring
3. Production management – physical distribution, facilities planning, manifestation maintenance and project scheduling
4. Marketing
5. Personnel management
6. Research and development.

3.3 Characteristics of Operations Research

Let us look at a few features here.

1. Interdisciplinary team approach in operations research, the optimisation solution was discovered by a team of scientists selected from various disciplines such as mathematics, statisticians, economics, engineering, physics etc.
2. Holistic approach to the system- operations research tries to find the best (optimum) decisions relative to largest possible portion of the total organisation.
3. Use of scientific research
4. Optimises the total output- operations research tries to optimise total return by maximising the profit and minimising cost or loss.

3.4 Phases of Operations Research

Operations research goes through a number of stages; let us take a look at the following.

1. Formulating the problem

To form an appropriate model, the following information will be required.

a. Who has to take the decision?
b. What are the objectives?
c. What are the ranges of controlled variables?
d. What are the uncontrolled variables that may affect the possible
e. solutions?
f. What are the restrictions or constraints on the variable?

Since wrong formulation cannot yield a right decision or solution. One must be considerably careful while executing these phases.

2. Constructing a mathematical model

This is the reformation of the problem in an appropriate form which is convenient for analysis. A mathematical model which represents the system under study is constructed for the purpose of solving the problem. A mathematical model must include:

a. decision variables and parameters
b. constraints or restriction
c. objective functions.

3. Deriving the solution from the model

This is devoted to the completion of those values of the decision variables that maximises or minimises the objective function. Such a solution is known as optimal solution, which is always appropriate for the problem under study.

4. Testing the model and its solution (updating the model)

A model is said to be valid if it can provide a reliable prediction of the system's performance. After completing the model, it is tested as a whole to find out errors if any. A model can be applicable for a longer time. It is necessary to update a model, from time to time, taking into account the past, present and future speculations on the problem.

5. Controlling the solution

The model requires immediate modification as soon as the controlled variables change significantly, if not the model goes out of control. As the conditions are constantly changing in the world, the model and the solution may not remain valid for a long time.

6. Implementing the solution

Finally, the tested results of the model are implemented to work.

SELF-ASSESSMENT EXERCISE

Critically analyse the merit and demerits of operations research.

4.0 CONCLUSION

In this unit, you have learnt what operations research is all about. You can now go on and apply the knowledge in your work.

5.0 SUMMARY

In this unit, you have learnt the essence of operations research as a new strategy for optimum development. You also learnt about the history of operations research, application of operations research and Phases of operations research include.

6.0 TUTOR-MARKED ASSIGNMENT

Discuss the phases of operations research.

7.0 REFERENCES/FURTHER READING

Kehine, J. (2008). *Quantitative Techniques and Statistical Methods*. Rakson Nigeria Limited.

Lucey, T. (1982). *Quantitative Techniques*. (6th ed.). British Publishers.

UNIT 2 LINEAR PROGRAMMING

CONTENTS

1.0 INTRODUCTION

In this unit, you will be exposed to linear programming- an element of operations research. It is used for minimising cost or minimising profit. Let us see how it works in terms of graphical approach.

2.0 OBJECTIVES

At the end of this unit, you should be available to:

- explain linear programming problems
- formulate linear programming problems, using graphical approach
- evaluate linear programming problems using graphical approach.

3.0 MAIN CONTENT

3.1 Linear Programming

Linear programming has to do with the linear function of variables called the objective function- subject to set of linear equations and / or inequalities called the constraints or restrictions. Some scientists, in 1947, while working in the *US* Air Force, observed that a large number of military programming and planning problems could be formulated as (minimising /maximising) a linear form of profit/cost function with variables restricted to values satisfying a system of linear constraints(a set of linear equations / or inequalities).

A linear form refers to a mathematical expression of the type- $a_1x_1 + a_2x_2 + \ldots a_nx_n$; where $a1,a_2,\ldots a_n$ are constants and $x_1,x_2\ldots x_n$ are variables. The term programming refers to the process of determining a particular program or plan of action.

3.2 Formulation of Linear Programming Problems

Here, you are to take note of the following.

1. Production allocation problem

A company produces two types of products *A* and *B*; it costs the company *N20* to produce *A,* and N24 to produce B. *A* is sold for N22, and *B* is sold for N21 (this means that the profits on *A* and *B* are *N2* and *N3,* respectively). Each product is processed on two machines- *G* and *H* . Product *A* requires one minute of processing time on *G,* and two minutes on type *H*; product **B** requires one minute of processing time on *G* and one minute on type *H*. The machine *G* is available for not more than 6 hours 40 minutes, while machine *H* is available for 10 hours, during any working day.

Solution
Let x1 be the number of products of product of type A
Let x2 be the number of products of product of type B

Table 2.1

Machine	Time of products (min)		Available time (min)
	TYPE A(X1 units)	TYPE B(X2 units)	
G	1	1	400
H	2	1	600
Profit per unit	N 2	N 3	

Objective function: maximise Z = 2x1 + 3x2

Subjected to*:*

1. machine G is not available for more than 6 hours 40 minutes, i.e. 400 minutes; therefore-
   $X_1 + X_2 \leq 400$ (first constraint)
2. also, machine H is available for 10 hours only, therefore-
   $2X_1 + X_2 \leq 600$ (second constraint)

3. since it is impossible to produce negative quantities
   $X_1 > 0, X_2 > 0$ (non-negativity constraints)

Hence, the allocation problem of the firm can be finally put in the form:

Maximise $Z = 2x1 + 3x2$

Subject to the conditions:

$X_1 + X_2 \leq 400$

$2X_1 + X_2 \leq 600$

$X_1 \geq 0, X_2 \geq 0$

Graphical Method

Simple linear programming problems of two decision variables can be easily solved by graphical method. The outlines of graphical procedure are as follows.

Step 1. Consider each inequality – constraint as equation
Step 2. Plot each equation on the graph, as each will, geographically, represent a straight line.
Step 3. Shade the feasible region. Every point on the line will satisfy the equation of the line. If the inequality – constraint corresponding to that line is <, then the region below the line lying in the first quadrant (due to non -0 negativity constraint of variables) is shaded.

The points lying in common region will satisfy the entire region.

Options are:

1. X1 = 300 x2 = 0
2. X1 =400 x2 = 0
3. X1 = 200 x2 = 200

Answer is option 2, with maximum value of **1200** i.e. produce 400 units of x1 and 0 unit of x2

4.0 CONCLUSION

You can now see that another approach to solving problems can be by using linear programming approach. All that you need to do is to formulate the problems; you can then proffer solution to the model you have set up.

5.0 SUMMARY

In this unit, you learnt that linear programming is a tool used in operations research to minimise problems. Also, it has been made clear to you that the basic steps to solving linear programming problems are as listed below.

- setup the objective function
- determine the constraints
- determine the break even line
- plot points
- evaluate - from the graph, the feasibility region that meets the goals and aspirations of the objective function.

6.0 TUTOR-MARKED ASSIGNMENT

i. Define the following terms - as used in linear programming:
   a. objective function
   b. constraints
   c. maximisation and minimisation in linear programming.

7.0 REFERENCES/FURTHER READING

Lucy, T. (1982). *Quantitative Techniques*. (6th ed.). British Publishers.

Oloyede, B. & Oluwakayode, E.F. (2001). *Quantitative Techniques for Business and Management Students*. *Volume 1*. Akure: Hope Printers and Publishers.

UNIT 3 TRANSPORTATION PROBLEM

CONTENTS

1.0 INTRODUCTION

In business (and life generally), material and human resources are, usually, moved from one place to another; and transporting these resources has cost implication. Every organisation wants to transport goods and services at a minimum cost. Transportation model is one of the techniques in operation research that is used in achieving this organisational objective.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- state the importance of transportation model in operation research
- explain how a transportation problem can be improved upon.

3.0 MAIN CONTENT

3.1 Transportation Problem

A typical transportation problem deals with sources where a supply of some commodity is available and destinations where the commodity is demanded.

The classic statement of the transportation problem uses a matrix with rows representing sources, and column representing destinations. The costs of transportation from the sources to destinations are indicated by

the entries in the matrix. Supply and demand are shown along the margins of the matrix.

A classic transportation problem has supply equal to total demand. Transportation problem is one of the subclasses of *LPP*- in which the objective is to transport various quantities of a single homogeneous commodity- that are initially stored at various origins, to different destinations in such a way that the total transportation cost is minimum.

3.2 Method of Solving Transportation Problem

Here, you will be exposed to the following:

1. north-west corner rule
2. least cost method
3. vogel approximation method.

1. North-West Corner (NWC) rule

Identify the cell at north corner of the matrix

a) Allocate as many units as possible to that cell, without exceeding supply or demand, and then cross out the row or column that is exhausted in doing this.
b) Reduce the amount of corresponding supply or demand (whichever is more) by the allocated amount.
c) Again identify the *NWC* cell of the reduced matrix
d) Repeat 2 and 3 until all the rows/columns requirements are satisfied.

2. Vogel Approximation Method (VAM)

a) Compute the difference between 2 smallest costs in the row or Column.
b) Identify the row/column with the largest difference. Find the first basic variable which has the smallest shipping cost in the row or column. Then, assign the highest possible value to that value and cross out the row/column which is exhausted.
c) Compute new differences and repeat the same procedure until all the rim requirements are satisfied.

3. Least Cost Method (LCM)

a) The lowest cost in the matrix is identified and units allocated from the source to the destination.

b) The next least cost is identified, and units allocated from the source to the destination.

Example

Nigeian brewery Plc is a company that has production operations in Lagos, Ibadan and Benin. The production capacity for each of these plants for the next 6 months planning period for Maltina is as follows.

Plant	**six month production capacity**
Lagos	5000
Ibadan	6000
Benin	2500
	13500

Suppose the company distributes through four regional distribution centres located in Sagamu, Warri, Kaduna and port-Harcourt and that the 6 months forcast demands for the disrtbution centres are as follows-

Distribution centre	forcast 6 – month Demand
Sagamu	6,000
Warri	4,000
Port Harcourt	2,000
Kaduna	1,500
	13,500

Management has compiled the transportation cost per unit from each of the plants to each of the distribution centres as shown in the table below:

	Distribution Centre			
To:	**Sagamu**	**Warri**	**P/Harcourt**	**Kaduna**
Lagos	3	2	7	6
from: Ibadan	7	5	2	3
Benin	2	5	4	5

Management would like to know how much of its production should be transported from each plant to each distribution centre.

Solution

Table3.1

To / From	**Sagamu**	**Warri**	**P/Harcourt**	**Kaduna**	**Supply Capacity**
Lagos	3 **5000**	2 **X**	7 **X**	6 **X**	5000
Ibadan	7 **1000**	5 **4000**	2 **1000**	3 **X**	6000
Benin	2 **X**	5 **X**	4 **1000**	5 **1,500**	2,500
Demand requirement	6,000	4000	2000	1500	13,500

Distribution schedule obtained using the North-West Corner(NWC) rule

From	To	Quantity	Unit cost	Total cost
Lagos	Sagamu	5000	3	15,000
Ibadan	Sagamu	1000	7	7,000
Ibadan	Warri	4000	5	20,000
Ibadan	P/Harcourt	1000	2	2,000
Benin	P/Harcourt	1000	4	4,000
Benin	Kaduna	1500	5	7,500
				55,500

Solution

Table 3.2

To / From	**Sagamu**	**Warri**	**P/Harcourt**	**Kaduna**	**Supply Capacity**
Lagos	3 **1000**	2 **4000**	7 **X**	6 **X**	5000
Ibadan	7 **2500**	5 **X**	2 **2000**	3 **1500**	6000
Benin	2 **2500**	5 **X**	4 **X**	3 **X**	2,500
Demand Requirement	6,000	4,000	2,000	1,500	13,500

1	(3)	2	2
1	x	2	2
(5)	x	2	2
(7)	x	2	3
X	x	2	(3)
X	x	(2)	x

Distribution schedule obtained by *VAM*

From	To	Quantity	Unit cost	Total cost
Lagos	Sagamu	1000	3	3,000
lagos	Warri	4000	2	8,000
Ibadan	Sagamu	2500	7	17,500
Ibadan	P/Harcourt	2000	2	4,000
Ibadan	Kaduna	1500	3	4,500
Benin	Sagamu	2500	2	5,000
				42,000

Evaluating initial feasible solution for possible improvement (optimality test).

The first task is to test for degeneracy, the number of used or occupied cells must be equal to $m+n-1$, where m is the number of rows- and n the number of columns. None of our initial solutions so far is degenerate.

Stepping stone method

Step 1- Select an unused cell to evaluate

Step 2- Beginning with the selected unused cell, trace a closed path, moving horizontally and vertically only, from the selected unused cell via used (occupied) cells back to the original unused cell. The path may run through unused cell being evaluated.

Step 3- Place alternative plus (+) and minus (-) signs at each corner of the closed path, starting with a plus sign at the unused cell being evaluated.

Step 4- Sum up the unit costs in the positive cells and deduct from it the sum of the unit costs ion the negative cells. The outcome is the improvement index #

Step 5- Repeat steps 1-4, until an improvement index has been obtained for each unused cell.

Decision criterion- if all the indices obtained are greater than or equal to zero, an optimal solution has been reached. If however, any of the indices is negative an improved solution is possible.

Improved solution

Step 6- Select the unused cell o route with the largest negative (in absolute term) improvement index.

Step 7- Identify the smaller allocation or assignment in a negative cell on the route.

Step 8- Add this figure (the smallest allocation) to all cells with plus signs and subtract it from cells with negative signs.

At the end of step 8, a new improved solution is obtained. We now test this new solution for optimality by repeating steps 1-5. If further improvement is possible, we proceed to steps 6-8. The process will continue until all indices are greater than or equal to zero.
We now test optimality some of the initial feasible solutions obtained by the North West-Corner (NWC) rule.

Rule and Vogel Approximation Method

Initial solution obtained by North-West Corner (NWC) rule

Solution

To / From	Sagamu	Warri	P/Harcourt	Kaduna	Supply Capacity
Lagos	(-) 3 **5000**	2 **X (+)**	7 **X**	6 **X**	5000
Ibadan	7 **1000 (+)**	5 **4000**	(-) 2 **1000**	3 **X (+)**	6000
Benin	2 **X**	5 **X**	4 **(+) 1000**	**5** **(-) 1,500**	2,500
Demand requirement	6,000	4000	2000	1500	13,500

There are six unused cells available for evaluation. Our objective is to find out what is the added or reduced cost of using any of these routes. To trace the closed path for Lagos-Warri route will involve the following movements:

Lagos – Warri	+2
Ibadan – Warri	-5
Ibadan – Sagamu	+7
Lagos – Sagamu	-3
	+1

Ibadan to Kaduna

Ibadan	– Kaduna	+3
Benin	– Kaduna	-5
Benin	– P/H	-5
Ibadan	– P/H	-5
		0

Lagos to Kaduna Route

To / From	Sagamu	Warri	P/Harcourt	Kaduna	Supply Capacity
Lagos	(-) 3 5000	2 X (+)	7 X	6 X	5000
Ibadan	7 1000 (+)	5 4000	(-) 2 1000	3 X (+)	6000
Benin	2 X	5 X	4 (+) 1000	5 (-) 1,500	2,500
Demand requirement	6,000	4000	2000	1500	13,500

Lagos – Kaduna	+6
Benin – Kaduna	-5
Benin – P/H	+4
Ibadan – P/H	-2
Ibadan –Sagamu	+7
Lagos – Sagamu	-3
	+7

Lagos to P/H Route

To / From	Sagamu	Warri	P/Harcourt	Kaduna	Supply Capacity
Lagos	(-) 3 **5000**	2 **X** **(+)**	7 **X**	6 **X**	5000
Ibadan	7 **1000** **(+)**	5 **4000** **(-)**	(-) 2 **1000**	3 **X** **(+)**	6000
Benin	2 **X**	5 **X**	4 **(+) 1000**	**5** **(-) 1,500**	2,500
Demand requirement	6,000	4000	2000	1500	13,500

Lagos – P/H	+7
Ibadan – P/H	-2
Ibadan – Sagamu	+7
Lagos – Sagamu	-3
	+9

Benin to Sagamu Route

To / From	Sagamu	Warri	P/Harcourt	Kaduna	Supply Capacity
Lagos	(-) 3 **5000**	2 **X** **(+)**	7 **X**	6 **X**	5000
Ibadan	7 **1000** **(+)**	5 **4000** **(-)**	(-) 2 **1000**	3 **X** **(+)**	6000
Benin	2 **X**	5 **X**	4 **(+) 1000**	**5** **(-) 1,500**	2,500
Demand requirement	6,000	4000	2000	1500	13,500

Benin – Sagamu	+2
Ibadan – Sagamu	-7
Ibadan - P/H	+2
Benin – P/H	-4
	-7

Benin to Warri Route

To / From	Sagamu	Warri	P/Harcourt	Kaduna	Supply Capacity
Lagos	(-) 3 **5000**	2 **X** **(+)**	7 **X**	6 **X**	5000
Ibadan	7 **1000** **(+)**	5 **4000**	(-) 2 **1000**	3 **X** **(+)**	6000
Benin	2 **X**	5 **X**	4 **(+) 1000**	**5** **(-)** **1,500**	2,500
Demand requirement	6,000	4000	2000	1500	13,500

Benin – Warri	+5
Ibadan – P/H	+2
Ibadan – Warri	-5
Benin – P/H	-4
	-2

At the end of the current exercise, we obtain the following result,

Route	Index
Lagos – Warri	+1
Ibadan – Kaduna	0
Lagos – Kaduna	+7
Lagos – Kaduna	+7
Benin – Sagamu	-7
Benin – Warri	-2

Improved solution is possible because two of the indices are negative. We now work on the route with the largest negat5ive, i.e. Benin – Sagamu Rout by performing step 6-8.

Improved solution

Benin – Sagamu Route

To / From	Sagamu	Warri	P/Harcourt	Kaduna	Supply Capacity
Lagos	3 **5000**	2 **X**	7 **X**	6 **X**	5000
Ibadan	7 **X**	5 **4000**	2 **2000**	3 **X**	6000
Benin	2 **1000**	5 **X**	4 **X**	5 **1,500**	2,500
Demand requirement	6,000	4000	2000	1500	13,500

Degeneracy

A careful study of the improved solution would show that the number of used cells are fewer than m + n + - 1. The imporoved solution degenerate and we shall not be able to trace a closed path for all the unused cells. To handle this situation, we artificially create a used cell, that is, we place a 0, representing nothing being transported in one of the unused cells and then treat it as if it were occupied. We must however choose the unused cell in which to place the 0 such that all stepping paths can be traced. We place the 0 at the Lagos – Warri cell.

Degeneracy problem solved

Benin – Sagamu Route

To / From	Sagamu	Warri	P/Harcourt	Kaduna	Supply Capacity
Lagos	3 **5000**	2 **0**	7 **X**	6 **X**	5000
Ibadan	7 **X**	5 **4000**	2 **2000**	3 **X**	6000
Benin	2 **1000**	5 **X**	4 **X**	5 **1,500**	2,500
Demand requirement	6,000	4000	2000	1500	13,500

The six cells to evaluate are-

Lagos-P/H, Lagos –Kaduna , Ibadan – Sagamu, Ibadan – Kaduna, Benin – Warri and Benin – P/H.

We proceed as we did previously. We shall spell out the cells involved in each path.

Lagos – Port Harcourt Route

Lagos – Port Harcourt	+7
Ibadan – Port Harcourt	-2
Ibadan – Warri	+5
Lagos – Warri	-2
	+8

Ibadan – Sagamu Route

Ibadan – Sagamu	+7
Ibadan – Warri	-5
Lagos – Warri	+2
Lagos - Sagamu	-3
	+1

Benin- Warri Route

Benin – Warri	+5
Lagos - Warri	-2
Lagos - Sagamu	+3
Benin - Sagamu	-2
	+4

Benin – Port Harcourt Route

Benin – Port Harcourt	+4
Ibadan – Port Harcourt	-2
Ibadan – Warri	+5
Lagos - Warri	-2
Lagos – Sagamu	+3
Benin – Sagamu	-2
	+6

Lagos - Kaduna Route

Lagos- Kaduna	+6
Benin – Kaduna	-5
Benin - Sagamu	+2
Lagos – Sagamu	-3
	0

Ibadan – Kaduna Route

Ibadan – Kaduna	+3
Benin – Kaduna	-5
Benin - Sagamu	+2
Lagos – Sagamu	-3
Lagos – Warri	+2
Ibadan – Warri	-5
	-6

Again, there is room for improved solution along Ibadan – Kaduna Route.

Improved solution: Ibadan – Kaduna

To / From	Sagamu	Warri	P/Harcourt	Kaduna	Supply Capacity
Lagos	3 **3500**	2 **1500**	7 **X**	6 **X**	5000
Ibadan	7 **X**	5 **2500**	2 **2000**	3 **1500**	6000
Benin	2 **2500**	5 **X**	4 **X**	5 **X**	2,500
Demand requirement	6,000	4000	2000	1500	13,500

The solution did not degenerate because the number of occupied cells equals *m+n-1*. We evaluate this solution for optimality.

Ibadan – Sagamu

Ibadan – Sagamu	+7
Ibadan – Warri	-5
Lagos – Warri	+2
Lagos – Sagamu	-3
	+1

Lagos- Port Harcourt

Lagos- Port Harcourt	+7
Ibadan – Port Harcourt	-2
Ibadan – Warri	+5
Lagos - Warri	-2
	+8

Lagos- Kaduna

Lagos- Kaduna	+6
Ibadan- Kaduna	-3
Ibadan – Warri	+5
Lagos - Warri	+2
	+6

Benin – Warri

Benin – Warri	+5
Lagos - Warri	-2
Lagos – Sagamu	+3
Benin – Sagamu	-2
	+4

Benin – Port Harcourt

Benin – Port Harcourt	+4
Ibadan – Port Harcourt	-2
Ibadan - Warri	+5
Lagos – Warri	-2
Lagos – Sagamu	+3
Benin – Sagamu	-2
	+6

Benin – Kaduna

Benin- Kaduna	+5
Ibadan – Kaduna	-3
Ibadan - Warri	+5
Lagos – Warri	-2
Lagos – Sagamu	+3
Benin – Sagamu	-2
	+6

Now we have all indices positive, there is no more room for improvement, our solution is optimal. The ₦39, 500 total transportation

cost is, at least, possible. It is lower than the total costs obtained in the initial solutions.

Optimal solution to the Nigerian Breweries' transportation problem

From	To	Quantity	Unit cost	Total cost
Lagos	Sagamu	3,500	3	10,000
Lagos	Warri	1,500	2	3,000
Ibadan	Warri	2,500	5	12,500
Ibadan	P/Harcourt	2,000	2	4,000
Ibadan	Kaduna	1,500	3	4,500
Benin	Sagamu	2,500	2	5,000
				39,500

3.5 Summary

In this unit, you learnt that:

(a) transportation problem is one of the subclasses of linear programming problem.
(b) the methods used for solving transportation problems are the North-west corner rule, least cost method and the Vogel approximation method.
(c) the objective of these techniques is to determine the maximum cost of transportation.

5.0 CONCLUSION

In this unit, it has been mentioned to you that transportation model is used to determine the minimum cost of transportation. The initial solution is improved upon, because of the assumption that there is no situation in the organisation that cannot be improved upon.

6.0 TUTOR-MARKED ASSIGNMENT

i. Consider the table below.

	W1	W2	W3	W4	Factory Capacity
F1	N19	N30	N50	N10	7
F2	N70	N30	N40	N60	9
F3	N40	N8	N70	N20	18
Ware House Requirments	5	8	7	14	34

Use the following methods to solve the transportation problem:

a. north-west corner rule
b. least cost method
c. Vogel approximation method

ii. Determine an initial basic feasible solution to the following transportation problem using the three methods.

a.

	D1	**D2**	**D3**	**D4**	**SUPPLY**
O1	6	4	1	5	**14**
O2	8	9	2	7	**16**
O3	4	3	6	2	**5**
DEMAND	**6**	**10**	**15**	**4**	**35**

b.

	I	**I**	**III**	**IV**	**SUPPLY**
A	13	11	15	20	**2000**
B	17	14	12	13	**6000**
C	18	18	15	12	**7000**
DEMAND	**3000**	**3000**	**4000**	**5000**	

7.0 REFERENCES/FURTHER READING

Olamade, O. (1998). *Problems and Solution in Production and Operations Research*. Wordsmiths Printing and Packaging Ltd.

Sharma, S. (2010). *Operations Research*: Theory, Methods and Applications. (15th ed.). India.

UNIT 4 GAMES THEORY

CONTENT

1.0 INTRODUCTION

In this unit, you will learn another approach to operation research; it is known as games theory. Here, choice of action is determined after taking into account all possible alternatives available to an opponent playing the same game, rather than just by the possibilities of several outcomes. The rule of thumb here is that, decision analysis is contingent upon maximum criterion.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- explain terms used in games theory
- describe principle of dominance.

3.0 MAIN CONTENT

3.1 Games Theory

Games theory, as mentioned earlier on, is a type of decision theory in which ones choice of action is determined, after taking into account all possible alternatives available to an opponent playing the same game, rather than just by the possibilities of several outcomes. The mathematical analysis of competitive problems is fundamentally based upon *minimax* (maximum) criterion. The criterion implies the assumption of rationality from which it is argued that each player will act so as to maximise his minimum gain or minimise his maximum loss.

Game is defined as an activity between two or more persons involving activities by each person according to a set of rules, at the end of which

each person receives some benefits or satisfaction, or suffers negative benefit.

3.2 Terms Used in Games Theory

1. Pay off- a quantitative or measure of satisfaction a person gets at the end of each play.
2. Strategy- a set of rules (programs) that specifies which of the available courses of action he should make at each play; one of the options a player can choose in a setting, when the outcome depends not only on his own actins but on the actions of others.
3. Two person zero sum game- a game with only two players if the losses of the player are equivalent to the gains of the other, so that the sum of the net gains is zero.
4. Saddle point- a saddle point of a pay off matrix is the position of such an element in the pay off matrix which is minimum in its row, and maximum in its column.

Rules for determining a saddle point.

a) Select the minimum element of each row of the payoff matrix and mark them by *0*
b) Select the greatest element of each column of the payoff matrix and mark with []

Strategy for player A	**1**	**2**	**3**	**4**	**5**	
	1	-2	6	0	5	3
Strategy for player B	**2**	3	2	1	2	2
	3	-4	-3	0	-2	6
	4	5	3	-4	2	-6

The best strategy for player A is strategy 2

The best strategy for player B is strategy 3

Saddle point is used to determine the optimal strategies players must use.

Note that when there is no saddle point, then optimal strategies are not pure strategies.

3.3 Principles of Dominance to Reduce the Size of a Game

If one pure strategy of a player is better or superior to another, then the inferior strategy may be simply ignored by assuming a zero probability, while searching for optimal strategies;i.e. if **a**= (a_1, a_2, a_3, a_n) and **b** =

(b_1, b_2, b_3, … b_n) and $a_i \geq b_i$ for all I = 1,2,3,….n, then **a** dominates **b.** The superior strategies are said to dominate the inferior ones.

Example

Player B	**I**	**II**	**III**	
PLAYER A	**I**	-4	6	3
	II	-3	-3	4
	III	2	-3	4

1. Row designations for each matrix are activities available to player A
2. Column designations for each matrix are activities available to player B
3. -4 is a loss to player A when be he plays strategy I and a gain of +4 to player B when he plays strategy I.
4. With a zero sum two person game, the cell entry in the player B's payoff matrix will be negative of the corresponding cell entry in the player A's payoff matrix so that the sum of payoff matrices for player A and player B is ultimately zero.

Solution

It is a clear that this game has a saddle point; strategy III is inferior to I for B. Strategy I dominates III, so therefore III is deleted for player B.

4.0 CONCLUSION

In this unit, you learnt the mathematical analysis of competitive problem which is based on *minimax*. You can now apply the strategy in your work. You can equally apply this as a management staff in you various offices.

5.0 SUMMARY

You learnt the following in this unit-

1. pay off- this is the satisfaction derived at the end of each play.
2. strategy- this is a set of programs that specifies which of the available courses of action a player should make at each play, depending on the action of other player(s).
3. Two Persons Zero Sum Game-The sum of the net gains is zero.

6.0 TUTOR-MARKED ASSIGNMENT

i. Define the following terms.
   a) Strategy
   b) Value of the game
   c) Saddle point
   d) Pay-off matrix

ii. Given the table below, find the saddle point.

Action	B1	B2	B3
A1	+6	+2	+4
A2	-6	0	+10

7.0 REFERENCES/FURTHER READING

Kehinde, J.S. (2008). *Quantitative Techniques and Statistical Method*. Rakson Nigeria Limited Educational Publisher.

Oloyede, .B. & Oluwakayode, E.F. (2001). *Quantitative Techniques for Business and Management Students*. *Volume 1*. Akure: Hope Printers and Publishers.

Sharma, S. (2010).*Operations Research: Theory, Methods and Applications*. (15th ed.). India.

UNIT 5 SIMULATION

CONTENTS

1.0 INTRODUCTION

In this unit, you will learn about simulation; this will assist you to solve complex operating system problems.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- define simulation
- state some application of simulation
- describe Monte-Carlo simulation model.

3.0 MAIN CONTENT

3.1 Simulation

Simulation has to do with the act of imitating the behaviour of situation or some process by means of something suitably analogous (especially for the purpose of study or personnel training). It is most often used whenever possible relationship cannot be expressed (as in the case of linear programming). It is a technique of dealing with dynamic system over times and it involves the manipulation of a model in order to ascertain the effect of likely changes in the value of the different variables that make up the system.

Many optimisation problems are difficult to formulate into known mathematical standard form for which solutions are available. Hence, simulation is used to address such a complex situation.

3.2 Areas of Application of Simulation

Simulation can be applied in the following areas-

1. Project schedule
2. Queuing modelling
3. Investing appraisal
4. Network analysis
5. Inventory control
6. Planning restaurant menus etc.

3.3 Stages in the Development of Simulation Model

Let us look at the following stages.

1. Determine the use of objectives- identify the reason why simulation model is required.
2. Identify thee input variables- i.e. controlled and non-controlled variable.
3. Identify the parameter- input variables which have constant values.
4. Determine probability distribution
5. Determine the logic of the model

3.4 Merits of Simulation

1. It provides solution where analytical methods are scarce.
2. There is less degree of assumption.
3. It can be applied even where the practices of the situation is not known.
4. It serves as a convenient management laboratory.

3.4.1 Demerits of Simulation

1. It does not necessarily produce optimal solution.
2. Simulation may lead to bias in statistical inference.
3. In project appraisal, simulation ignores the benefit of diversification in risk reduction.
4. The financial model required running the simulation is often very complex and might be too expensive to construct.

3.5 Types of Simulation Models

1. Deterministic Model- this a large scale model that can be used to ask "what if" type of question- e.g. what is the percentage drop in volume of sale?
2. Probabilistic model- this is known as Monte-Carlo or Bootstrapping method.

3.5.1 Steps Involved in Carrying-Out Monte-Carlo Simulation

To use Monte-Carlo simulation model, two tables are usually constructed, taking the following steps:

1. construction of a cumulative column
2. assigning random number range to each possible outcome on the basis of cumulative probability.
3. Simulation table:

    a. Write down the numbers of run serially and vertically.
    b. Determine the forecast level of the variable by using random numbers allocation.
    c. If no monetary consideration is attached, sum the value of forecast demand and obtain the mean etc. Mean equals the total forecast values dividend by the number of runs.

Example 1

Salvation Enterprises is seeking advice on the level of a certain material to carry on daily basis, so as to maximise its profit. From experience, the record of past sales of material is given as follows.

Table 5.1

POSSIBLE DEMAND	PROBABILITY
5,000	0.17
8,000	0.13
9,500	0.12
10,000	0.19
10,500	0.04
11,000	0.35

Attempt a simulation and advice Salvation Enterprises on the optimal quantity to carry, using the following random numbers.

50, 53, 22, 54, 96, 65, 24, 14, 57, 73, 54, 77, 69 and 52.

Solution

Table 5.2

Possible Demand	**Probability**	**Cumulative Probability**	**Random Number**
5,000	0.17	0.17	00-16
8,000	0.13	0.30	17-29
9500	0.12	0.42	30-41
10,000	0.19	0.61	42-60
10,500	0.04	0.65	61-60
11,000	0.35	1.00	65-99
	1.00		

Table 5.3

RUNS	**RANDOM NUMBER**	**FORECAST DEMAND**
1	50	10,000
2.	53	10,000
3.	22	8,000
4.	54	10,000
5.	96	11,000
6.	95	11,000
7.	65	11,000
8.	24	8,000
9.	14	5,000
10.	57	10,000
11.	73	11,000
12.	54	10,000
13.	77	11,000
14.	69	11,000
15.	52	10,000
	Total	**147,000**

Recommendation

The optimal quantity $= \frac{147,000}{15}$

$= \mathbf{9,800}$

Therefore, Salvation Enterprises should carry 9,800 units of the material for maximum profit.

4.0 CONCLUSION

This unit will assist you to deal with a complex problem solving situation. You can apply it in investment appraisal, inventory control and evaluating queuing system. There are soft-wares that will assist you to generate iterations of simulated numbers. You need to purchase some of them.

5.0 SUMMARY

In this unit, you have learnt some areas where simulation can be used. You have also learnt the stages in the development of simulation and how to use random number generation to run simulation model. I will encourage you to apply what you have learnt in your work.

6.0 TUTOR-MARKED ASSIGNMENT

i. Discuss the different between Monte-Carlo Simulation and bootstrapping.

ii. Eternity Ventures maintain a current account with Jesus Bank plc with special provision for automatic overdraft facility. If the account balance falls below N11, 000, 000.00, the bank automatically transfers ₦ 8, 500, 000.00 into the account as loan. The probability distribution of the company's daily net deposit and withdrawal are as follows.

Daily Deposit		**Daily Deposit**	
Amount (₦)	**Probability**	**Amount (₦)**	**Probability**
4,500,000	0.27	5,000,000	0.17
5,000,000	0.11	5,500,000	0.25
5,500,000	0.07	6,000,000	0.16
6,000,000	0.34	6,500,000	0.14
6,500,000	0.05	7,000,000	0.05
7,000,000	0.16	7,500,000	0.23

The random numbers in respect of deposit and withdrawals are as follows-

Deposit: 21, 49, 63, 04, 16, 55, 79, 01, 83 and 42.
Withdrawal: 72, 45, 17, 48, 40, 22, 75, 69, 86 and 45
Required:

(a) simulate for 10 days
(b) determine the average account balance

(c) determine the average time between loan transfers into the account.

Note: assume that the initial balance is (₦)5, 000, 000.00

7.0 REFERENCES/FURTHER READING

Kehinde, J.S. (2008). *Quantitative Techniques and Statistical Method*. Rakson Nigeria Limited Educational Publisher.

Oloyede, .B. & Oluwakayode, E.F. (2001). *Quantitative Techniques for Business and Management Students*. *Volume 1*. Akure: Hope Printers and Publishers.

UNIT 6 NETWORK ANALYSIS

CONTENTS

1.0 INTRODUCTION

There is the need to have a work-plan for activities when a project is about to be executed. Each activity should be known and arranged in sequential order. The order of arrangement should be based on the nature of each activity. Planning these different activities are part of the feasibility studies necessary for project execution in any organisation.

2.0 OBJECTIVES

At the end of this unit, you should be able to:

- determine the critical plan in a network diagram
- illustrate the probability of completing a project within a given period.

3.0 MAIN CONTENT

3.1 Terms Used in Project Management

A project is a specific work or assignment with a starting point and an ending point. It is a series of related jobs or tasks focused on the achievement of an overall objective. A project contains various

networks of activities- one preceding the other. It consists of a series of network of activities that is meant to complete a job.

Project management involves planning, controlling and directing resources (people, equipment and materials) to meet the technical, cost and time constraints of the project.

Network analysis provides the methodology for planning and controlling projects- e.g. the introduction of a new product. The main aim of network analysis is to monitor the progress of a project so that the project is completed within the minimum possible time. In doing this, the technique pin-points the parts of a project which are critical, i.e. those parts which, if delayed beyond the allotted time, would delay the completion of the entire project. Let us look at some terms common with project management.

1. **Activity** - a certain amount of work required in the project.
2. **Critical activity**-an amount that has no room for schedule slippage; if it slips, the entire project completion will slip. It is an activity with zero slack.
3. **Dummy activity**- an activity that consumes no time but shows precedence among activities.
4. **Critical path**- the chain of critical activities in a project. The longest path through the network. It determines the length of a project.
5. **Events**- this is the beginning, completion point within the project. An activity begins and ends with events. It is usually represented by a circle- "0".
6. **Slacks**- the amount of time that an activity or a group of activities can slip without causing a delay in the completion of the project.
7. **Predecessor activity**- an activity that must occur before another.
8. **Successor activity**- an activity that must occur after another activity.
9. **Earliest finish**- the earliest time that an activity can finish from the beginning of a project.
10. **Earliest start**- the earliest that an activity can start from the beginning of the project.
11. **Latest finish**- the latest that an activity can finish from the beginning of the project without causing a delay in the completion of the project.
12. **Latest start**- the latest that an activity can start from the beginning of the project without causing a delay in the completion of the project.

A network diagram representation is as given below.

Start event *activity* **End event**

Activity	A	B	C	D	E	F	G	H	I	J	K
Preceding activity	-	A	A	C	B,C	D,E	E	G	D,F	I,H	J

3.2 Goals of project network analysis

1. To estimate completion time of a project and possibly reduce such completion time at a reasonable cost afforded by the project manager.
2. Critically analyses the process and various activities to be carried out with the specific time allotted for completion.
3. To reduce the total cost of the project.

3.3 Methods Used in Solving Project Network Analysis Problems (Project Management Techniques)

1. Critical Path Method (CPM)
2. Project Evaluation and Review Technique (PERT)

PERT and *CPM* are protect management techniques which have been created out of the need of the western industrial and military establishments to plan, schedule and control complex projects. Planning, scheduling (or organising) and controlling are considered to be basic managerial functions, and *CPM/PERT* has been rightfully accorded important in operations research. *PERT/CPM* provide a focus around which managers can brainstorm and put their ideas together. It becomes a useful tool for evaluating the performance of individuals and teams.

These techniques can answer the following important questions:

1. how long will then entire project take to be completed? What are the risks involved?
2. what are the critical activities or tasks in the project ?

3.3.1 Basic Steps in PERT/CPM Techniques

1. Planning

Splitting the total project into small projects; these smaller projects, in turn, are divided into activities and are analysed by the department or section. Relationships between activities are defined.

2. Scheduling

A time chart showing the start and finish times for each activity, as well .as its relationship to other activities. This pin-points the critical path activities which require special attention, if the project is to be completed in time.

3. Allocation of resources

A resource is a physical variable such as labour, finance, equipment and space which will impose a limitation on time for the project. When resources are limited and conflicting demands are made for the same type of resources, a systematic method for allocating resources becomes crucial.

4. Controlling

This is the final stage of *PM*; having progress report from time to time and updating the network continuously, a better financial and technical control over the project is also of crucial importance. The two stages that form the basis for project control are as listed below.

- **Preparing the network**

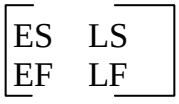
  This involves:

  a. defining the objective of the project
  b. identifying the individual jobs which make up the project
  c. determining the logical sequence of jobs.
  d. determine the estimated duration for each activity
  e. construct the appropriate diagram.

- **Analysis of the network**

  a. Determine the critical path
  b. Determine the project completion time/ duration
  c. Determine the slack times

     i. Find the *ES* and *EF* for each activity
     ii. Find the *LS* and *LF* for each activity
     iii. Determine the total slack time for each activity.



Slack = LF – EF
Slack = LS – ES

3.4 Float (Slack)

Float is defined as a measure of delay it is the amount of time that a task in a project network can be delayed without causing a delay to subsequent tasks (free float), project completion time (total float). Jobs which are critical have no float. The most reliable method for determine the critical path and establishing the delay on no-critical jobs is based on a calculation of float.

3.4.1 Classification of Float

1. Total float
2. Free float
3. Independent float

1. Total float

The amounts of time by which an activity can be delayed without affecting the completion time of the project. Total float is associated with the path. Total float represents the schedule flexibility and can be measured by subtracting early dates from late dates of completion time.

2. Free float

The period for which an activity can be delayed, without affecting subsequent activities.

3. Independent float

Amount of time by which an activity can be delayed, if preceding activities are completed -as late as possible and subsequent activities are started as early as possible.

NOTE: An activity on critical path has zero free float, but an activity may not be on the critical path. Total float is associated with the path. If a project network chart/diagram has 4 non-critical paths then that project will have *4* total float values. The total float of a path is the combined free float values of all activities in a path.

SELF-ASSESSMENT EXERCISE

What is network analysis?

3.5 The Critical Path Method (CPM)

CPM is used when definite activity time and events in accomplishing a project is known. The technique is applied in two stages- to prepare the network and analyse the network.

Example1

Activity	**Designation**	**Imm. Predecessor**	**Time in days**
Design	A	-	21
Build Prototype	B	A	4
Evaluate Equipment	C	A	7
Test Prototype	D	B	2
Write Equipment Report	E	C,D	5
Write Methods Report	F	C,D	8
Write Final Report	G	E,F	2

Required-

1. Construct the network diagram
2. Determine the critical path
3. Determine the project completion time
4. Determine the total slack time for each activity.

Solution

1. Path**: A-C-F-G = 38 days**
   Path : A-C-E-G = 35 days
   Path : A-B-D-F-G = 37 days
   Path : A-B-D-E-G = 34
2. Path A-C-F-G takes the longest period of time to complete- which is 38 days and is therefore the critical path for this project.

3. Total slack for each activity

Activity	LS	ES	SLACK (LS – ES)	LF	EF	SLACK (LF - EF)
A	0, ~~1~~	0	0	21	21	0
B	22	21	1	26	25	1
C	~~24~~,21	21	0	28	28	0
D	26,~~29~~	25	1	28	27	1
E	31	~~27~~,28	3	36	33	3
F	28	~~27~~,28	0	36	36	0
G	36	36,~~33~~	0	38	38	0

3.5.1 How to Compute LS, ES, LF, AND EF

1. Earliest start

This is determined by a forward pass, through the network, beginning with the first activity- i.e. for which we set *ES = 0* representing the start of the project. To find ES_B -add the time it takes to complete activity *A* (21 days) to the duration of *B.*

When more than one activity precedes the activity being evaluated, then the *ES* for each path leading to that activity is calculated, the latest *ES* is selected -e.g. *ES* for Activity *F*,

ES_F = Max (ES_C + C, ES_D + D)
ES_F = Max (28,27)
ES_F = 28.

2. Latest start

A backward pass through the network, stating from the end of the project- i.e. activity G,
LS for activity G
LS_G = Completion time – G
LS_G = 38-2 = 36

To find LS of F

$LS_F = LS_G – F$ =36 – 8 = 28

When more than one activity follow the activity, being evaluated, then the *LS* for all paths leading out of the activity must be completed and that path with the earliest *LS* is used-

i.e. LS_C = Min ($LS_F – C$, $LS_E – C$)
LS_C = Min (28-7, 31-7)
LS_C = Min (21,24) = 21 days

3. Earliest finish time

Begin at activity *A* and move from left to right *forward pass* across the network to determine the *EF* value for each activity. For all activities that begin a project, their *EFs* are their durations e.g. ES_A = 21. For all other activities, an activity's *EF* is the *EF* of its immediate predecessor activity plus its duration,i.e.- $EF_B = EF_A + B = 21 + 4 = 25$

When an activity has two or more immediate predecessor activities, the longest *EF* among the immediate predecessor activities must be used in computing its *EF*.

4. Latest finish time

Begin from the end of the project i.e. *G* and move from right to left, *backward pass,* through the network. *LF* for all activities that end in the last event will always be the greatest EF of the project. The *LF* for any other activity is computed by subtracting the immediate successor activity's duration (the activity to its immediate right in the network) from the immediate successor activity's *LF*.
When an activity has more than one immediate succeeding activity, its *LF* is computed by comparing the values of *LF – D* for all the immediate succeeding activities. The smallest *LF- D* value is used as its *LF*.

SELF-ASSESSMENT EXERCISE

Activity	A	B	C	D	E	F	G	H	I
Imm predecessor	-	A	B	A	C,D	E	D	E	G,H
Duration	20	10	8	11	7	6	12	13	5

Required-

i. Daw the network diagram
ii. Determine the critical path
iii. Determine the project completion time
iv. Determine slacks for each activity.

3.6 Project Evaluation and Review Technique (PERT)

PERT estimates three possible time in accomplishing a project network, the variable of interest is time. The three-time estimate used are:

i. optimistic time (a) if all goes well
ii. normal time (b) best estimate
iii. pessimistic time (c) in case of unforeseen contingencies.

Expected time, $ET = (a + 4b + c)/6$
Variance $\delta^2 = ((c - a)/6)^2$
Standard deviation $\delta = \sqrt{\delta^2}$

3.6.1 Procedure for Solving PERT

Take note of the following.

1. Identify each activity to be done in the project.
2. Determine the sequence of activities and construct a network reflecting the precedent relationship.
3. Determine the three time estimates for each activity
4. Calculate the expected time (ET) for each activity
5. Calculate the variance (σ^2) of each activity.
6. Identify all the path in the network and their estimated completion time
7. Calculate variances and standard deviation of the critical path
8. Determine the probability of completing the project by a given date.

Example

Activity	Immediate predecessor	Time estimates in weeks		
		A	B	C
A	-	10	22	28
B	A	1	4	7
C	A	4	6	14
D	B	1	2	3
E	C,D	1	5	9
F	C,D	7	8	9
G	E,F	2	2	2

Required:

i. Draw the network diagram and identify the critical path.
ii. Calculate the expected time *ET* for each activity and the completion duration of the project.
iii. Calculate the variance for each activity.
iv. What is the probability of completing the project in 39 days?

Solution

1.

Activity	Imm Pred	Time Estimates in Weeks			Expected Time	Activity Variance
		A	B	c		
A	-	10	22	28	21	9.0
B	A	1	4	7	4	1.0
C	A	4	6	14	7	2.78
D	B	1	2	3	2	0.11
E	C,D	1	5	9	5	1.78
F	C,D	7	8	9	8	0.11
G	E,F	2	2	2	2	0

2. path**: A-C-F-G = 38 days**
   Path : A-C-E-G = 35 days
   Path : A-B-D-F-G = 37 days
   Path : A-B-D-E-G = 34 days
3. critical Path = A-C-F-G
4. completion duration = 38 weeks
5. variance of the critical path- δ^2= 9 + 2.78 + 0.11 + 0 =11.89 weeks
6. standard deviation of the critical path- $\delta_p = \sqrt{11.89} = 3.45$ weeks
7. the probability of completing the project in 39 weeks:
   here we use the normal distribution table. To do this *Z* must be computed -
   i.e. $Z = (D - E_p)/\delta_p$
   Where D = Desired completion time
   E_p = Expected completion time
   δ_p = Standard deviation of the critical path
   $Z = (39 - 38)/3.45 = 0.289$.

In the table locate Z ≈ 0.29, the answer is 0.6141.

The probability that the project will be completed in less than 39 weeks is 61.41%

Note- the probability that the project will take longer than 39 weeks is 100 - 61.41 = 38.59%.

4.0 SUMMARY

In this unit, you learnt that a project consists of the period of network of activities that is meant to complete a job. The largest path through the network is the critical path. There are two major methods used in solving project network; these are *CPM* and *PERT*.

5.0 CONCLUSION

In this unit, it has been made clear to you that network analysis is an important project management technique used for achieving organisational objectives. It is widely used in engineering and construction companies.

6.0 TUTOR-MARKED ASSIGNMENT

i. State the basic rules for drawing networks.

Job No	**Pred. Job**	**Time Estimate**		
		a	**b**	**c**
1	**-**	2	3	4
2	**1**	1	2	3
3	**1**	4	5	12
4	**1**	3	4	11
5	**2**	2	3	5
6	**3**	1	2	3
7	**4**	1	8	9
8	**5,6**	2	4	6
9	**8**	2	4	12
10	**7**	3	4	5
11	**9,10**	5	7	8

Required

i. Construct the appropriate network diagram.
ii. Identify the critical path.
iii. What is the expected completion time of the project?
iv. What is the probability that the project will be completed in 30 days?
v. Calculate total, independent and free float for all the activities.

7.0 REFERENCES/FURTHER READING

Chandara, P. (2009). *Project Analysis, Selection, Financing, Implementation and Review.* (7th ed.) New Delhi: McGraw, Hill Publishing Company Ltd.

Daris, M., Aquilano, N. & Chase, R. (2003). *Fundamentals of Operation Management*. (5th ed.). New York: McGraw - Hill.

Sharma, S. (2010). *Operations Research: Theory, Methods and Applications*. (15th ed.). India.