NATIONAL OPEN UNIVERSITY OF NIGERIA

SCHOOL OF MANAGEMENT SCIENCES

COURSE CODE: BHM 722

COURSE TITLE: BUSINESS STATISTICS

NATIONAL OPEN UNIVERSITY OF NIGERIA

SCHOOL OF MANAGEMENT SCIENCES

COURSE GUIDE

Course Code: **BHM 722**

Course Title: **BUSINESS STATISTICS**

Course Developer/Writers: **Mr. KADIRI Kayode I. (National Open University of Nigeria)**

Mr. SUFIAN Jelili B. (National Open University of Nigeria)

Editor:

Programme Leader: **Dr. C.I. Okeke (National Open University of Nigeria)**

Course Coordinator: **Mrs. IHUOMA Ikemba-Efughi (National Open University of Nigeria)**

CONTENTS

INTRODUCTION:

Business Statistics is a one semester, 3 credit units first year level course. It will be available to all first degree of the school of Business and Human resources Management at the National Open University, Nigeria. It will also be useful for those seeking introductory knowledge in business statistics.

The course consists of eighteen units that involved basic concepts and principles of statistics and decision making process, forms of data, methods of data collection, summarizing data, graphical presentation of data, measures of both central tendency and dispersion, set theory, permutations and combinations, some elements of probability concepts, probability distributions of both discrete and continuous random variables.

The course requires you to study the course materials carefully, supplement the

materials with other resources from Statistics Textbooks both to be prescribed and those not prescribed that may treat the contents of the course.

This Course Guide tells you what the course is about, what course materials you will be using and how you can work your way through these materials. It suggests some general guidelines for the amount of time you are likely to spend on each unit of the course in order to complete it successfully. It also gives you some guidance on your tutor--marked assignments. Detailed information on tutor-marked assignment is found in the separate file.

There is likely going to be regular tutorial classes that are linked to the course. It is advised that you should attend these sessions. Details of the time and locations of tutorials will be communicated to you by National Open University of Nigeria (NOUN).

What You Will Learn In The Course

The overall aim of BHM 722 Business Statistics is to introduce you to the nature of Statistical Information, Collection, Summarizing and Presentation and analyzing the statistical information in such a way that the reality contained in the information may be revealed for decision-making. During the course, you will be exposed to nature of statistical information, collection and processing of statistical data. You will also be exposed to elementary concepts in probability and nature, characteristics and uses of some important probability distributions of both discrete random variable and continuous random variable with research method.

Course Aims

The course aims to give you an understanding of statistical information and presentation for decision-making. It exposes you to measures that are computed and used for processing materials for decision-making. It also gives the basic knowledge of some concepts used for making decisions and carefully summarizes some Probability Distributions.

This will be achieved by:

1. Introducing you to nature and form of statistical data
2. Showing how the statistical data can be collected and presented
3. Showing you how to compute measurement of dispersion in a sample or population
4. Showing you how to compute value of permutations and combinations
5. Introducing you to the basic concepts of elementary probability
6. Give the basic principles for the application of some important probability distributions

Course Objectives

To achieve the aims set above the course sets overall objectives; in addition, each unit also has specific objectives. The unit objectives are included at the beginning of a unit; you should read them before you start working through the unit. You may want to refer to them during your study of the unit to check on your progress. You should always look at the unit objectives after completing a unit. In this way you can be sure you have done what was required of you by the unit.

We set out wider objectives of the course as a whole below. By meeting these objectives, you should have achieved the aims of the course.

On successful completion of the course, you should be able to:

1. Explain nature of Statistical information.
2. Explain types of Statistical information
3. Collect Statistical information
4. Summarize Statistical information
5. Present Statistical information
6. Compute measures of Central Tendency and Dispersion for Statistical information
7. Explain basic concepts in set theory
8. Perform operations in set theory
9. Compute the values of Permutations and combinations for arrangement of objects
10. Define Probability of an event
11. Explain Properties of Probability
12. Calculate Probability events
13. Explain concepts used in probability
14. Explain the principles underlying the application of various probability distributions
15. Compute measures for Probability Distributions.
16. Explain the uses of T-test, F-test
17. Compute the value of chi-square
18. Know the uses of ANOVA

Working through This Course

To complete this course, you are required to read the study units, read set books and other materials on the course.

Each unit contains self-assessment exercises called Student Assessment Exercises, SAE. At some points in the course, you are required to submit assignments for assessment purposes. At the end of the course there is a final Examination. This course should take about 22 weeks to complete. Some listed components of the

course, what you have to do and how you should allocate your time to each unit in order to complete the course successfully on time, are given below

Below you will find listed components of the course, what you have to do and how you should allocate your time to each unit in order to complete the course successfully on time.

Course Materials

Major components of the course are:

(1) Course Guide
(2) Study Units
(3) Textbooks
(4) Assignment File
(5) Presentation Schedule.

Study Units

The course is in five modules and twenty-four Study Units as

follows:

MODULE ONE:

Unit 1: Statistics and Decision Making Process
Unit 2: Nature, Source and Method of Data Collection
Unit 3: Summarizing Data
Unit 4: Graphical Presentation of Data
Unit 5: Research Methods

MODULE TWO:
Unit 1: Measures of Central Tendency 1- The Arithmetic Mean
Unit 2: Measures of Central Tendency 2 - Geometric Mean and Harmonic Mean.
Unit 3: Measures of Central Tendency 3 - Median and Mode
Unit 4: Measures of Dispersion

MODULE THREE:
Unit 1: Set theory
Unit 2: Permutations and Combinations
Unit 3: Some Elementary Probability Concepts
Unit 4: Probability Rules, Events and Bayes' Theorem
Unit 5: Probability Distribution of a Discrete Random Variable

MODULE FOUR:
Unit 1: Correlation Theory
Unit 2: Pearson's Correlation Coefficient
Unit 3: Spearman's Regression Analysis
Unit 4: Lease Square Regression Analysis
Unit 5: Multiple Regression Analysis

MODULE FIVE:
Unit 1: T - t e s t
Unit 2: F - t e s t
Unit 3: C h i - s q u a r e
Unit 4: ANOVA

The first five units concentrate on the nature, collection and presentation of statistical data. This constitutes Module 1. The next five units, module 2, concentrate on computation of measures of central tendency and dispersion from samples and populations. Module3, deal with the basic concepts and principles in elementary probability. Module 4, teach the principles underlying the applications of some important probability distributions. The last five units, Module 5, teach the principles underlying the applications of some important test of statistical theory.

Each unit consists of one week direction for study, reading material, other resources and summaries of key issues and ideas. The units direct you to work on exercises related to the required readings

Each unit contains a number of self-tests. In general, these self-tests question you on the material you have just covered or required you to apply it in some way and thereby help you to assess your progress and to reinforce your understating of the material. Together with tutor-marked assignments, these exercises will assist you in achieving the stated learning objectives of the individual units and of the course.

Set Textbooks

It is advisable you have some of the following books

JUDE I.E, MICAN & EDIITH Statistics& Quantitative Methods for Construction & Business Managers.
OKOJIE, Daniel E. NOUN Statistics for Economist.

AJAYI J. NOUN Business Statistics 1.

Assignment File
In this tile, you will find the details of the work you must submit to your tutor for marking. The marks you obtain for these assignments will count toward the final mark you obtain for this course. Further information on assignments will be found in the Assignment File itself and later in this Course Guide in the section on Assessment.

There are four assignments in this course. The four course assignments will cover:

Assignment 1 - All question in Units 1 - 5
Assignment 2 - All TMAs' question in Units 1 - 4
Assignment 3 - All TMAs' question in Units 1 - 5
Assignment 4 - All TMAs' question in Units 1 - 4.
Assignment 5 - All TMAs' question in Units 1 - 4.

Presentation Schedule
The presentation schedule included in your course materials gives you the important dates for this year for the completion of tutor-marking assignments and attending tutorials. Remember, you are required to submit all your assignments by due date. You should guide against falling behind in your work.

Assessment

There are two types of the assessment of the course. First are the tutor-marked assignments; second, there is a written examination.

In tackling the assignments, you are expected to apply information, knowledge and techniques gathered during the course. The assignments must be submitted to your tutor for formal Assessment in accordance with the deadlines stated in the Presentation Schedule and the Assignments File. The work you submit to your tutor for assessment will count for 50 % of your total course mark.

At the end of the course, you will need to sit for a final written examination of three hours' duration. This examination will also count for 50% of your total course mark.

Tutor-Marked Assignments (TMAs)

There are five tutor-marked assignments in this course. You are advice to exercise all the assignments. You are encouraged to work all the questions thoroughly. Each assignment counts toward your total course mark.

Assignment questions for the units in this course are contained in the Assignment File. You will be able to complete your assignments from the information and

materials contained in your set books, reading and study units. However it is desirable in all degree level education to demonstrate that you have read and researched more widely than the required minimum. You should use other references to have a broad viewpoint of the subject and also to give you a deeper understanding of the subject.

When you have completed each assignment, send it, together with a TMA form, to your tutor. Make sure that each assignment reaches your tutor on or before the deadline given in the Presentation File. If for any reason, you cannot complete your work on time, contact your tutor before the assignment is due to discuss the possibility of an extension. Extensions will not be granted after the due date unless there are exceptional circumstances.

Final Examination and Grading

The final examination will be of 2½ hours' duration and have a value of 50% of the total course grade. The examination will consist of questions which reflect the types of self testing, practice exercises and tutor-marked problems you have previously encountered. All areas of the course will be assessed

Use the time between finishing the last unit and sitting the examination to revise the entire course. You might find it useful to review your self-tests, tutor-marked assignments and comments on them before the examination. The final examination covers information from all parts of the course.

COURSE OVERVIEW

This table brings the units together, the number of weeks you should take to complete them and the assignment as follows.

UNIT	TITLE OF WORK	WEEKS ACTIVITY	ASSESSMENT (END OF UNIT)
	COURSE GUIDE		
	MODULE ONE: STATISTICS MEANING & PRESENTATION		
1	Statistics& Decision Making	1	Assignment 1
2	Nature, Source & method of Data collection	1	Assignment 2
3	Summarizing Data	1	Assignment 3
4	Graphical Presentation of Data	1	Assignment 4
5	Research Methods	1	Assignment 5
	MODULE TWO: CENTRAL TENDENCY		
1	Measurement of Central Tendency 1	1	Assignment 6
2	Measurement of Central Tendency 2	1	Assignment 7
3	Measurement of Central Tendency 3	1	Assignment 8
4	Measurement of Dispersion	1	Assignment 9
	MODULE THREE: SET THEOREM & PROBABILITY		
1	Set Theory	1	Assignment 10
2	Permutations & Combinations	1	Assignment 11
3	Some Elementary Probability Concepts	1	Assignment 12
4	Probability Rules, Events & Bayes' Theorem	1	Assignment 13
5	Probability Distribution	1	Assignment 14
	MODULE FOUR: CORRELATION THEORY AND REGRESSION ANALYSIS		
1	Correlation Theory	1	Assignment 15
2	Pear's Correlation Coefficient	1	Assignment 16
3	Spearman's Correlation Coefficient	1	Assignment 17
4	Lease Square Regression Analysis	1	Assignment 18
5	Multiple Regression Analysis	1	Assignment 19
	MODULE FIVE: STATISTICAL TEST		
1	T-test	1	Assignment 20
2	F-test	1	Assignment 21

3	Chi-Square	1	Assignment 22
4	ANOVA	1	Assignment 22

How to Get the Most from This Course

In distance learning the study units replace the university lecturer. This is one of the great advantages of distance learning; you can read and work through specially designed study materials at your own pace and at a time and place that suit you best. Think of it as reading the lecture instead of listening to a lecturer. In the same way that a lecturer might set you some reading to do, the study units tell you when to read your books or other material, and when to undertake computing practical work. Just as a lecturer might give you an in-class exercise, your study units provides exercises for you to do at appropriate points.

Each of the study units follows a common format. The first item is an introduction to the subject matter of the unit and how a particular unit is integrated with the other units and the course as a whole. Next is a set of learning objectives. These objectives let you know what you should be able to do by the time you have completed the unit. You should use these objectives to guide your study. When you have finished the unit you must go back and check whether you have achieved the objectives. If you make a habit of doing this you will significantly improve your chances of passing the course.

The main body of the unit guides you through the required reading from other sources. This will usually be either from your set books or from a Readings section. Some units require you to undertake practical work or a computer. You will be directed when you need to use a computer and guided through the tasks you must do. The purpose of the computing work is twofold. First, it will enhance your understanding of the material in the unit. Second, it will give you practical experience of using programs, which you could well encounter in your work outside your studies. In any event, most of the techniques you will study are applicable on computers in normal working practice, so it is important that you encounter them during your studies.

Self-tests are interspersed throughout the units, and answers are given at the ends of the units. Working through these tests will help you to achieve the objectives of the unit and prepare you for the assignments and the examination. You should do each self-test as you come to it in the study unit. There will also be numerous examples given in the study units; work through these when you come to them, too.

The following is a practical strategy for working through the course. If you run into any trouble, contact your tutor. Remember that your tutor's job is to help you. When you need help, don't hesitate to call and ask your tutor to provide it.

1. Read this Course Guide thoroughly.

2. Organize a study schedule. Refer to the `Course overview' for more details. Note the time you are expected to spend on each unit and how the

assignments relate to the units. Important information, e.g. details of your tutorials, and the date of the first day of the semester is available from study centre. You need to gather together all this information in one place, such as your dairy or a wall calendar. Whatever method you choose to use, you should decide on and write in your own dates for working breach unit.

3. Once you have created your own study schedule, do everything you can to stick to it. The major reason that students fail is that they get behind with their course work. If you get into difficulties with your schedule, please let your tutor know before it is too late for help.
4. Turn to Unit 1 and read the introduction and the objectives for the unit.

5. Assemble the study materials. Information about what you need for a unit is given in the `Overview' at the beginning of each unit. You will also need both the study unit you are working on and one of your set books on your desk at the same time.

6. Work through the unit. The content of the unit itself has been arranged to provide a sequence for you to follow. As you work through the unit you will be instructed to read sections from your set books or other articles. Use the unit to guide your reading.
7. Up-to-date course information will be continuously delivered to you at the study centre.

8. Well before the relevant due date (about 4 weeks before due dates), get the Assignment File for the next required assignment. Keep in mind that you will learn a lot by doing the assignments carefully. They have been designed to help you meet the objectives of the course and, therefore, will help you pass the exam. Submit all assignments no later than the due date.

9. Review the objectives for each study unit to confirm that you have achieved them. If you feel unsure about any of the objectives, review the study material or consult your tutor.

10. When you are confident that you have achieved a unit's objectives, you can then start on the next unit. Proceed unit by unit through the course and try to pace your study so that you keep yourself on schedule.

11. When you have submitted an assignment to your tutor for marking do not wait for it return `before starting on the next units. Keep to your schedule. When the assignment is returned, pay particular attention to your tutor's comments, both on the tutor-marked assignment form and also written on the assignment. Consult your tutor as soon as possible if you have any questions or problems.

12. After completing the last unit, review the course and prepare yourself for the final examination. Check that you have achieved the unit objectives (listed at

the beginning of each unit) and the course objectives (listed in this Course Guide).

Tutors and Tutorials

There are some hours of tutorials (ten 2-hour sessions) provided in support of this course. You will be notified of the dates, times and location of these tutorials. Together with the name and phone number of your tutor, as soon as you are allocated a tutorial group.

Your tutor will mark and comment on your assignments, keep a close watch on your progress and on any difficulties you might encounter, and provide assistance to you during the course. You must mail your tutor-marked assignments to your tutor well before the due date (at least two working days are required). They will be marked by your tutor and returned to you as soon as possible.

Do not hesitate to contact your tutor by telephone, e-mail, or discussion board if you need help. The following might be circumstances in which you would find help necessary. Contact your tutor if.

- You do not understand any part of the study units or the assigned readings
- You have difficulty with the self-tests or exercises
- You have a question or problem with an assignment, with your tutor's comments on an assignment or with the grading of an assignment.

You should try your best to attend the tutorials. This is the only chance to have face to face contact with your tutor and to ask questions which are answered instantly. You can raise any problem encountered in the course of your study. To gain the maximum benefit from course tutorials, prepare a question list before attending them. You will learn a lot from participating in discussions actively.

Summary

Business Statistics I intend to introduce you to the nature and types of statistical Data, Collection, Presentation and analysis of the data to bring out the realityof the information contained in the data for decision making. Upon completion of this course, you will be equipped with the basic knowledge in data and analyzing data to obtain the reality of the information contained in them. You will be able to answer these kind of question:

- What are the natures of statistical data?
- What are the methods of collecting Statistical data?
- How can statistical data be presented?
- The important measures computed from statistical data?
- How are these measures computed?

- What are the concepts in elementary Probability?
- What are the characteristics of probability?
- How can you compute the probability of an event?
- What are the unique characteristics of probability distributions?
- How can you use the research Methods to solve problems

What are the concepts in research methods?

How can you use Hypothesis in research Methods to solve problems

There are more and more questions you can ask. To gain a lot from the course, please try to apply anything you learn in the course to real life situations. We wish you success with the course and hope that you will find it both interesting and useful.

NATIONAL OPEN UNIVERSITY OF NIGERIA

SCHOOL OF MANAGEMENT SCIENCES

Course Code: **BHM 722**

Course Title: **BUSINESS STATISTICS**

Course Developer/Writers: **Mr. KADIRI Kayode I. (NOUN)**

Mr. SUFIAN Jelili B. (NOUN)

Editor:

Programme Leader: **Dr. C.I. Okeke (NOUN)**

MODULE ONE: STATISTICAL MEANING AND PRESENTATION

UNIT 1:
STATISTICS AND DECISION MAKING PROCESS

CONTENTS

1.0 Introduction

You will realize that the activities of man and those of the various organizations, that will often be referred to as firms, continue to increase. This brings an increase in the need for man and the firms to make decisions on all these activities. The need for the quality and the quantity of the information required to make the decisions increases also. The management of any firm requires scientific methods to collect and analyze the mass of information it collects to make decisions on a number of issues. Such issues include the sales over a period of time, the production cost and the expected net profit. In this regard, statistics plays an important role as a management tool for making decisions.

2.0 Objectives

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

- Understand the various definitions of statistics
- Describe the uses of statistics
- Define the basic concepts in statistics.

3.1 Definitions of Statistics

Statistics can be defined as a management tool for making decision. It is also a scientific approach to presentation of numerical information in such a way that one will have a maximum understanding of the reality represented by such

information. Statistics is also defined as the presentation of facts in numerical forms. A more comprehensive definition of statistics shows statistics as a scientific method which is used for collecting, summarizing, classifying, analyzing and presenting information in such a way that we can have thorough understanding of the reality the information represents.

From all these definitions, you will realize that statistics are concerned with numerical data.. Examples of such numerical data are the heights and weights of pupils in a primary school when evaluating the nutritional well being of the pupils and the accident fatalities on a particular road for a period of time.

You should also know that when there are numerical data, there must be non-numerical data such as the taste of brands of biscuits, the greenness of some vegetables and the texture of some joints of a wholesale cut of meat. Non-numerical data cannot be subjected to statistical analysis except they are transformed to numerical data. To transform greenness of vegetables to numerical data, a five point scale for measuring the colour can be developed with 1 indicating very dull and 5 indicating very green.

3.2 The Roles of Statistics

You will realize that statistics is useful in all spheres of human life. A woman with a given amount of money, going to the market to purchase foodstuff for the family, takes decision on the types of food items to purchase, the quantity and the quality of the items to maximize the satisfaction she will derive from the purchase. For all these decisions, the woman makes use of statistics

Government uses statistics as a tool for collecting data on economic aggregates such as national income, savings, consumption and gross national product. Government also uses statistics to measure the effects of external factors on its policies and to assess the trends in the economy so that it can plan future policies.

Government uses statistics during census. The various forms sent by the government to individuals and firms on annual income, tax returns, prices, costs, output and wage rates generate a lot of statistical data for the use of the government

Business uses statistics to monitor the various changes in the national economy for the various budget decisions. Business makes use of statistics in production, marketing, administration and in personnel management.

Statistics is also used extensively to control and analyze stock level such as minimum, maximum and reorder levels. It is used by business in market research to determine the acceptability of a product that will be demanded at various prices by a given population in a geographical area. Management also uses statistics to make forecast about the sales and labour cost of a firm. Management uses statistics to establish mathematical relationship between two or more variables for the purpose of predicting a variable in terms of others. For the conduct and analyses of biological, physical, medical and social researches, we use statistics extensively.

3.3 Basic Concepts in Statistics

Let us quickly define some of the basic concepts you will continue to come across in this course.

- **Entity**: This may be person, place, and thing on which we make observations. In studying the nutritional well being of pupils in a primary school, the entity is a pupil in the school.

- **Variable**: This is a characteristic that assumes different values for different entities. The weights of pupils in the primary school constitute a variable.

- **Random Variable:** If we can specify, for a given variable, a mathematical expression called a function, which gives the relative frequency of occurrence of the values that the variable can assume, the function is called a probability function and the variable a random variable.

- **Quantitative Variable:** This is a variable whose values are given as numerical quantities.

  Examples of this is the hourly patronage of a restaurant

- **Qualitative Variable:** This is a variable that is not measurable in numerical form or that cannot be counted. Examples of this are colours of fiuits, taste of some brands of a biscuit.

- **Discrete Variable:** This is the variable that can only assume whole numbers. Examples of these are the number of Local Government Council Areas of the States in Nigeria, number of female students in the various programmes in the National Open University.

  A discrete variable has "interruptions" between the values it can assume. For instance between 1 and 2, there are infinite number of

values such as 1.1, 1.11, 1.111, 1. I I l land so on. These are called interruptions.

- **Continuous Variable:** This is a variable that can assume both decimal and non decimal values. There is always a continuum of values that the continuous variable can assume. The interruptions that characterize the discrete variable are absent in the continuous variable. The weight can be both whole values or decimal values such as 20 kilograms and 220.1752 kilograms.

- **Population:** This is the largest number of entities in a study. In the study of how workers in Nigeria spend their leisure hours, the number of workers in Nigeria constitutes the population of the study.

- **Sample:** This is the part of the population that is selected for a study. In studying the income distribution of students in the National Open University, the incomes of 1000 students selected for the study, from the population of all the students in the Open University will constitute the sample of the study.

- **Random Sample:** This is a sample drawn from a population in such a way that the results of its analysis may be used to generalize about the population from which it was drawn.

Exercise 1.1

What is the importance of Statistics to human activities? Your answer can be obtained in section 3.2 of this unit.

4.0 Conclusion

In this unit you have learned a number of important issues that relate to the meaning and roles of statistics. The various definitions and examples of concepts given in this unit will assist tremendously in the studying of the units to follow.

5.0 Summary

What you have learned in this unit concerns the meaning and roles of statistics, and the various concepts that are important to the study of statistics.

6.0 Tutor Marked Assignment

What is Statistics? Of what importance is statistics?

7.0 References, Further Reading and Other Sources

Daniel, W.W. and Terrel J.C. (1979) Business Statistics: Basic Concepts And Methodology 2"a ed. Houghton Mifflin Company Boston.

Hannapan, T.J. (1982) Mastering Statistics. The Macmillan Press Ltd.

UNIT 2:
STATISTICS AND DECISION MAKING PROCESS

CONTENTS

1.0 Introduction

This is the second unit in the course. The first unit has shown that statistics involves the scientific method of collecting, summarizing, classifying, analyzing and presenting information, usually numerical information, in a way that we can have a thorough understanding of the reality the information represents. You will realize that it is important to know the nature, the sources and methods of collecting the data. This unit takes you through these aspects of data. There are self-assessment exercises designed to make you pause and reflect on what you are reading.

At the end of the unit, there are again self-assessment questions which are designed to know the extent of your learning of the contents in the unit.

2.0 Objectives

After studying unit and going through the exercises, you should be able to

- Know the nature of primary and secondary data
- Identify various sources of micro-and macro-statistical information
- Describe the various methods of data collection

3.1 Nature of Statistical Data

3.1.1 Primary Data

You are already aware that statistics uses numerical data. The Numerical data can be divided into:

a) Primary and
b) Secondary Data

Primary Data are data collected by or on behalf of the person or people who are going to make use of the data. It is the data collected specifically for a purpose and used for the purpose for which they are collected.

Examples of Primary data are:

(i) Heights and weights of students collected to determine their nutritional well being.
(ii) The population of Primary school pupils in the states of the country to allow the Federal Government plan for the primary education.
(ii) The academic performance of students in secondary schools under various types of leaders to know the leadership style suitable for the secondary schools

The various methods of collecting primary data include surveys, interview, observation, questionnaire and experiments. These will be further discussed in this unit.

Primary data could be very expensive to collect when the elements of the study sample are widely scattered and when the items of equipment for data collection, as in many experiments, are capital intensive. However errors can be minimized when collecting primary data since the researcher can always take adequate precaution in collecting primary data.

3.1.2 Secondary Data

This is the data that is used by a person or people other than the person or people by whom or for whom the data was collected. These are the data collected for some other purpose, frequently for administrative reasons, and used for the purpose for which they were not collected.

Secondary data are always collected from published sources, like textbooks, journals, Newspapers, magazines, and gazette.

Examples of secondary data include:

(i) Accident fatalities on a particular road over a period of time collected from the Police or Road safety corps.
(ii) Dietary requirements of various age groups collected from a nutrition textbook.
(iii) Age distribution in Nigeria collected from the publications of the National Population Census.

From the discussion so far, you will know that secondary data are second-hand data. This shows the need to know as much as possible about the data. In trying

to know much about the secondary data, we need to consider the following points:

(i) How the data was collected
(ii) How the data has been processed
(iii) The accuracy of the data
(iv) How far the data has been summarized
(v) How comparable the data is with other tabulations.
(vi) How to interpret the data, especially when figures collected for one purpose are used for another.

With secondary data, we usually strike a compromise between what we want and what we are able to find.

Secondary data have the following advantages:

(i) They are very inexpensive to collect since they are readily and abundantly available in published sources.
(ii) There is a great variety of secondary data on a wide range of subjects.
(iii) Many of these data have been collected for many years hence we can use them to establish trends for forecasting.

With all these advantages of secondary data, we must use secondary data with great care since such data may not give the exact kind of information required and the data may not be in the most suitable form they are required.

Student Assessment Exercise 2.1

For each of the Primary and Secondary data list:

(i) The characteristics
(ii) The advantages
(iii) The disadvantages

3.2 SOURCES OF DATA

3.2.1 Micro-Statistical Information

The firms and private organizations, when monitoring the activities of their businesses, produce a lot of information that are specific to the organization and firm and used for their decision-making processes. The information produced by these form and organization are called micro-statistical Information.

In micro-economics, the concern is for the household and firm. The information produced or generated at these levels are micro-statistical information. For firm, such information include those generated from production, marketing and personnel.

3.2.2 Macro-Statistical Information

These are produced by the Public sector of the economy. They are related to the whole country as a whole. Such information include population, education, rate of inflation, level of unemployment and so on.

Macro-economics is concerned with such aggregates as national income, gross national products, saving, consumption, gross domestic product and so on. These are macro-statistics information.

3.2.3 Statistical Information in Business

In The performance and monitoring of various activities in the firms, a lot of data is produced. The quality and the quantity of the information generated in the firm depend on the size of the firm and the resources available to the firm

The firm is interested in what is happening to the national economy and what is happening in the industry it belongs to. You should know that a combination of firms makes the industry. As regards national income, the firm wishes to know the interest rates and unemployment. As regards the industry, the firm will always wish to know wage rates, prices, level of output so that it can compare its performance with the competitors in the industry.

The firm also generates data in the following areas.

(i) Production:

The firm collects data on the results of quality control, prices of raw materials, defectiveness of consignment of raw materials, wage rates, stock of materials, accident rates, absenteeism rates, unit cost.

(ii) Marketing:
The firm will to want know the budgeted sales, the advertising cost to sustain the sales, the distribution costs etc.

Other information could be generated from personnel and Accounts departments to aid decision-making process of the firm.

3.2.4 Government Statistics

As you were informed in unit 1 section 3.2, the governments produce statistics to be able to measure the effects of their policies, to monitor the effects of external factors on their policies and to be able to assess trends so that they can plan future policies. The macrostatistical information is generated by the governments as you have learned in this unit. The various governments rely in the firms and the individuals to generate these macro statistical data

Student Assessment Exercises 2.2

a) Look at a firm around you and list ten pieces of information the firm generates from the activities of the firm.
c) Look at your Local Government Council Area list ten pieces of Information the council generates for the governance of the council.

3.3 Methods Of Data Collection

3.3.1 Surveys

In this unit, you have learned that secondary data are already available; hence they must be collected when there is the need to use them. The collection of primary data involves survey or inquiry of one type or the other.
Some surveys can be limited in the sense that they can be carried out with a few minutes of observation. Others can be detailed. When surveys are detailed, information from the surveys are more acceptable and valued than when they are limited.

Examples of surveys are:

(i) Government Survey

(ii) Market research surveys - carried out for one particular client and not published in any form.
(iii) Research surveys - carried out by academicians and published in journals
(iv) Firms commission ad hoc surveys on a wide variety of subjects.

Survey methods consist of the following stages:

(i) The survey design - This depends on the objectives of the survey, the available methods, the amount of money, and time that can be allocated for collecting the information
(ii) The Pilot survey: This is the preliminary survey carried out on a very small scale to make sure that the design and methodology of the survey are likely to produce the information required

(iii) Collection of Information - This involves the use of observation, interview and questionnaire.
(iv) Coding - We may need to pre-code the questions to facilitate classification and tabulation.
(v) Tabulation - There is the need to tabulate the data, to give a summary of the data.
(vi) Secondary statistics - We will need to calculate secondary statistics such as means and percentages.
(vii) Reports - Reports must be written on the results and the results must be illustrated with graphs and diagrams.

3.3.2 Observations

This is one of the methods of collecting primary data. It can be used to know the use a particular facility is put. It can be used to study the behaviours of people in a work place.

There can be the following types of observation:

- Participant observation
- Systematic observation
- Mechanical observation

In participant observation, the observer is involved in the activities he is trying to observe. Examples of these are the vice-chancellor who participates in the eating at the cafeteria to observe the performance of the cooks and the acceptability of the food by the students; the lecturer who sits for his own paper with the students to observe the difficulty encountered by the students in the paper. This method can have serious influence in the entities the observer is observing. The method may also consume a lot of time.

In systematic observation, the observer does not take part in the activity. The method is used when events can be investigated without the participants knowing that somebody is observing them. Though the method is objective, it does not question the motives of people observed.

In mechanical observation, mechanical devices do the observation. For instance, the number of vehicles passing a particular point on the road can be recorded mechanically. Sophisticated mechanical means such as television, film and tape recorders are used to provide more complex information. Mechanical observations can be more effective than those observation made by individuals observer who can be subject to bias.

Generally, observation has the following problems:

- Objectivity- to remain objective, the observer cannot ask the question that will help him to understand the events he is observing.
- Selectivity- an observer can unintentionally become selective in perception, recording and reporting.
- Interpretation- the observer may impute meanings to the behaviour of people that the people do not intend.
- Chance- a chance event may be mistaken for a recurrent one.
- Participation- the participation of the observer can influence the behaviour of people being observed.

3.3.3 Interviewing

This is a conversation with a purpose. There can be formal and informal interviews. Informally everybody uses interviews to obtain information.
The formal interview is also initiated by the interviewer who approaches the person he is interested in interviewing. The interviewer therefore arranges the venue, and the time, and prepares the questions to be asked. The interviewer also secures the means of recording the responses.

Interviews are used in a number of situations that include:

(i) Obtaining opinion polls for the success of a candidate in an election.
(ii) Studying why people behave in a way
(iii) Selecting applicants for some jobs
(iii) Reporting especially in the radio and television

Before the interview is conducted, the interviewer must have knowledge of his respondents. The interviewer must also secure an initial rapport with his respondents. The interviewer must explain clearly and briefly the purpose of the interview to his respondents.

The interviewer should be very objective. He must not express his own opinions and must not influence the answers of the respondents. The language of the question must be at the level of the respondents. If the questions are written in a language different from that of the respondents, the question must be translated to that of the respondent but the answers must be written by the interviewer in the language of the questions. This was the case during the last census in Nigeria. The questionnaire used was written in English language. That is not the language many Nigerians understand. For the purpose of getting the response of people in this group; the questions were translated to the language they understand.

Interviewing method has a number of advantages:

(i) It allows the interviewer to have personal contact with the respondents therefore allowing more questions to be asked which improves the quality of the information.
(ii) It allows the interviewer to persuade unwilling respondents.
(iv) It allows experienced interviewer to know when to make calls and recalls

Interviewing method also has a number of disadvantages.

(i) Biased interviewer may influence the responses of the respondents to suit his own opinions.
(ii) A biased respondent, in a matter-affecting ego, may give false responses.
(iii) It is very expensive and time consuming especially when respondents are widely scattered.
(v) It may be difficult to interview some top people in government and business

3.3.4 Questionnaire

This is a list of questions drawn in such a way that the questions are related to the objectives of the study being conducted, and the responses to the question will be analyzed to provide solutions to the problems we attempt to solve in the study.

There are two types of questionnaire namely:

(i) Structured or fixed-response questionnaire:
(ii) Unstructured or open ended questionnaire

The structured questionnaire consists of a list of questions drawn on the study being conducted. Each question is accompanied by alternative answers from which the respondent picks appropriate answer or answers. An example of a structured question is this:

What is your monthly salary

- Below N7500
- N7500 - N10,000.0

0

- N10, 000 - N12, 500.0
0
- Above N12,500

Structured questionnaire has a number of advantages:

(i) It is very easy to complete and analyze.
(u) Most of the questions are answered
(iii) The responses are always related to the objectives of the study.

A major disadvantage of the structured questionnaire is the fact that it does not allow the views of the respondents, which may enhance the quality of the information collected.

Unstructured questionnaire is a list of questions drawn on the study on which information is required. The questions are not accompanied by alternative answers as in the structured questionnaire. The respondents are free to provide their own responses.
Example of a question in an unstructured questionnaire is "What is your monthly salary?" Unstructured questionnaire are not difficult to construct since no question is accompanied by alternative answers. It has the following advantages.

(i) It allows for the views of the respondents
(ii) The questionnaire are not difficult to construct since no question is accompanied by alternative answers.

However, the unstructured questionnaire has the following disadvantages.

(i) It is not easy to complete and analyze.
(ii) Many of the responses supplied by the respondents may not be related to the objectives of the study.
(iii) The respondents may not answer all questions, especially those considered difficult by them. This will definitely reduce the quality of the information obtained.

In drawing a standard questionnaire, the following must be considered.

(1) List all the questions you want to ask.
(ii) Decide whether to use direct or indirect questions or both
. (iii) A question must be limited to an idea.
(v) Ask simple and interesting questions before the difficult and uninteresting ones.
(v) State the questions clearly.

(vi) A question must mean the same thing to all the respondents

(vii) The language must be at the level of the respondents.
(viii) Do not ask questions that will hurt your respondent
(ix) The questionnaire should be pre-tested on a mock-audience before it is administered on the real sample to detect the difficulty in completing and analyzing the questionnaire so as to review the questionnaire before it finally gets to the real audience it is meant for.

Student Assessment Exercise 2.3

You may wish to put off your course material, pause for a while and list the various methods of data collection, their descriptions, their advantages and disadvantages

4.0 Conclusion

In this unit, you have learned primary data and secondary data; their nature, examples, sources, advantages and disadvantages. You should also be aware of the various methods of collecting data, their advantages and disadvantages. In the unit to follow, you will learn how to summarize data to facilitate the necessary analysis that will bring out the reality contained in the information.

5.0 Summary

What you have learnt in this unit consists of the nature of statistical data, the sources of the data and the methods of collecting the data. The unit to follow will build upon this.

6.0 Tutor Marked Assignment

What are the distinguishing features between interviews and observation?

7.0 References, Further Reading And Other Resources

Ajayi, J.K. (1997) Elements of Business Statistics. Unpublished Monograph, Ondo State Polytechnic8, Owo

Hannagan, T.J. (1982) Mastering Statistics. The Macmillan Press Ltd. Pg. 33-42

UNIT 3:
SUMMARIZING DATA

CONTENTS

1.0 Introduction

You learned the methods of collecting data in the previous unit. In the first unit, you have learned that statistics is a scientific method of collecting, summarizing, classifying, analyzing and presenting information in such a way that we can have a thorough understanding of the reality the information represents. From this definition of statistics, summarizing the data is very important to the analysis of the data. You will learn in this unit how to arrange the mass of unordered array of mass of data in such a way that detailed analysis on the data can be performed.

2.0 Objectives

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

- Arrange an unordered array of data in ordered form.
- Prepare frequency distributions for both grouped and ungrouped data.
- Construct relative frequency, cummulative relative frequency for both ungrouped and grouped data

3.1 Ordered Array

When data are collected, they are collected in such a way that there is no particular arrangement of the values. This unordered array of values does not facilitate the process of analyzing the data. The data must therefore be arranged either in ascending or descending order of magnitude so as to facilitate the analysis.

Example 3.1: Suppose the ages of twenty pupils in a primary school are as follows (age to the nearest years)

6, 8, 10, 11, 7, 5, 8, 9, 11, 12. 11, 9, 8, 12, 5, 6, 10, 8, 7, 9.

A look through the values shows no order of arrangement. An ordered form of the values shows the following:

5, 5, 6, 6, 7, 7, 8, 8, 8, 8. 9,9,9,10,10,11,11,11,12,12.

The data arranged in this form has a number of advantages over the raw data.

(i) We can quickly know the lowest and highest values in the data.
(ii) We can easily divide the data into sections.
(iii) We can see whether any value appears more than once in the array.
(iv) We can observe the distance between succeeding values in the data.

Student Assessments Exercise 3.1

Go through the values listed below and arrange them in ascending order of magnitude

The daily births in a maternity home in a month are given below

5, 3, 6, 7, 4, 1, 2, 4, 0, 3
3, 5, 2, 6, 2, 5, 0, 7, 2, 5
2, 3, 5, 4, 6, 7, 5, 1, 1, 3

3.2 Frequency Distribution

3.2.1 Frequency Distribution for Ungrouped Data

An ordered array of data does not sufficiently summarize the data. The data can be further summarized by preparing the frequency distribution for the data.

A frequency distribution is a table showing the values of the data and the number of occurrence of each of the values.

Example 3.2

For the data in exercise 3.1 above, present the frequency distribution for the values. The values will be represented by Xi and the frequency is represented by Fi.

Frequency Distribution

Xi	Fi
0	2
1	3
2	5
3	5
4	3
5	6
6	3
7	3

This table shows the values of the variable and their respective frequencies.

3.2.2 Frequency Distribution of Grouped Data

It is possible to have the frequency in example 3.2 without grouping the data because the values are not many. There are only seven values. In situation where the values are in thousands or millions, it may be difficult to analyze the values if they are not grouped.
For many of sophisticated analyses, we need to group the data before analysis commences. We therefore group the data into class intervals

Class intervals: are defined as contiguous, non-overlapping intervals selected in such a way that they are mutually exclusive and collectively exhaustive. They are mutually exclusive in the sense that a value is Placed in one and only one class interval.
The class intervals could be 5-9, 10-14, 15-19, 20-24..... It could also be 11-20, 21-30, 31-40,.....
It may even take either form. In this unit there will be more examples of class intervals. The class should not be too few and should not be too many. Too few class intervals can result in a loss of much detail while too many class intervals may not condense the

Example 3.2

The table below shows the scores of 50 students in mathematics in a Senior Secondary Examination

19	50	57	25	61	42	26	33	46	45
63	31	80	36	78	56	38	69	83	40
52	17	35	65	13	63	72	29	56	57
22	45	53	44	76	47	86	55	66	48
41	64	38	43	23	58	55	32	52	46

For this we need to prepare the tallies from the tallies we obtain the frequency of each of the class intervals.

Class Intervals	Tallies	Frequency
11-20	III	3
21-30	~~IIII~~	5
31-40	IIII III	8
41-50	IIII IIII I	11
51-60	~~IIII~~ ~~IIII~~	10
61-70	~~IIII~~ II	7
71-80	IIII	4
81-90	II	2

What you have above is the frequency distribution of a grouped data. To further summarize grouped data, some basic concepts are important. These concepts will be defined and computed now.

(i) **Class Limit:** For any class interval, there are two class limits, the lower and upper class limits. For the example 3.2, the lower class limits are 11, 21, 31, 41, 51, 61, 71, 81. The upper class limits are 20, 30, 40,50, 60, 70, 80 and 90.

(ii) **Class Boundaries**: For any class interval, there are two class boundaries. The lower class boundary of a class interval is the mean of

the lower class limit of the interval and the upper limit of the preceeding interval. Let us compute the lower class boundary of the interval 11-20. The lower class limit of the class interval is 11 and the upper class limit of the preceeding class interval is suppose to be 10. The lower class boundary of the class interval is therefore:

$$\frac{10+11}{2} = 10.5$$

For the example 3.2, the lower class boundaries of the class intervals are 10.5,20.5, 30.5, 40.5, 50.5, 60.5, 70.5, and 80.5.

The upper class boundary of a class interval is the mean of the upper class limit of the class interval and the lower class limit of the succeeding class interval. For example 3.2, the upper class boundary of interval 11-20 is 20.5. The upper class limit of the class interval is 20 while the lower class limit of the succeeding class interval is 21. The upper class boundary of the interval is therefore equal to:

$$\frac{20+21}{2} = 20.5$$

For the example, the upper class boundaries of the class intervals are respectively 20.5, 30.5, 40.5, 50.5, 60.5, 70.5, 80.5, and 90.5. From these values, you will see that the upper class boundary of a class interval is the lower class boundary of the succeeding class interval. The classes can also be given in terms of class boundaries rather than class limits. When this is done, there is overlapping of the class intervals. For example 3.2, if the class boundaries are used, we will have the following as our frequency distribution.

(iii) **Class Width:** This is the difference between the upper and lower

Class Intervals	Frequency
10.5-20.5	3
20.5-30.5	5
30.5-40.5	8
40.5-50.5	11
50.5-60.5	10
60.5-70.5	7
70.5-80.5	4
80.5-90.5	2

class boundaries (not class limits). For our example 3.2, the class width for all the class intervals is 10. For the first interval, the class width is 20.5-10.5 = 10. It is not 20-11.

(iv) Class Mark: This is the mean of the upper and the lower class boundaries. It can also be the mean of the lower and the upper class limits. For example 3.2, the class mark for the first interval is:

$\frac{10.5+20.5}{2} = 15.5$. It can also be $\frac{11+20}{2} = 15.5$

For the classes in the example 3.2, we have the following as class width for the respective class intervals 15.5, 25.5, 35.5 , 45.5, 55.5, 65.5, 75.5, and 85.5. There is need to summarize what we have done so far into class limits, class boundaries, class marks and class width.

Class Limits	Class Boundaries	Class Marks	Class Width
11-20	10.5-20.5	15.5	10
21-30	20.5-30.5	25.5	10
31-40	30.5-40.5	35.5	10
41-50	40.5-50.5	45.5	10
51-60	50.5-60.5	55.5	10
61-70	60.5-70.5	65.5	10
71-80	70.5-80.5	75.5	10

81-90	80.5-90.5	85.5	10

3.3 Relative Frequency

3.31 Ungrouped Data

The relative frequency of a value is defined by the total frequencies of all the values contained in the set of values.

Example 3.3. Suppose the frequency distribution of the scores of the twenty students in a test is as presented below

Scores	Frequency
2	1
3	2
4	2
5	4
6	5
7	3
8	2
9	1
TOTAL	20

The total frequency is 20. The relative frequency of the first score is given as $1/20 = 0.05$

The relative frequency distribution of the scores is therefore given as:

Scores	Frequency	Relative Frequency
2	1	0.05
3	2	0.10
4	2	0.10
5	4	0.20
6	5	0.25
7	3	0.15
8	2	0.10
9	1	0.05

3.3.2 Grouped Data

For the grouped data, the relative frequency of a class interval is the frequency of the interval divided by the total frequencies of all class intervals. For example 3.2, the relative frequency distribution of the class intervals is as presented

Relative Frequency Distribution of Grouped Data

Class Intervals	Frequency	Relative Frequency
11-20	3	0.06
21-30	5	0.10
31-40	8	0.16
41-50	11	0.22
51-60	10	0.20
61-70	7	0.14
71-80	4	0.08
81-90	2	0.04

Going through the relative frequencies table presented in the unit, you will realize that the sum of the relative frequencies for a set of values is 1

3.4 Cumulative Relative Frequency Distribution

3.4.1 Ungrouped Data

In this unit, you have learned construction of relative frequency for ungrouped data. A further summary of data can be in form of cumulative relative frequency. To construct the cumulative relative frequency, we need to construct the cumulative frequency for the data. The cumulative relative frequency of a value is the cumulative frequency of the values divided by the total frequency of the values contained in the array of data.

For the example 3.3, the cumulative frequency and the cumulative relative frequency for the set of data is as presented below

Score	Frequency	Cumulative Frequency	Cumulative Relative Frequency
2	1	1	1/20 = 0.05
3	2	3 = 1+2	3/20 = 0.15
4	2	5 = 3+2	5/20 = 0.25
5	4	9= =5+4	9/20 = 0.45
6	5	14=9+5	14/20=0.70
7	3	17 = 14+3	17/20 =.85
8	2	19 = 17+2	19/20 =.095
9	1	20 = 19+1	20/20 = 1.00

3.4.2 GROUPED DATA

The cumulative frequency of grouped data is constructed the same way as that of ungrouped data.

Example 3.4

The table below shows the frequency distribution of the ages of employees in a firm. Prepare the cumulative frequency and the cumulative relative frequency for the ages of the employees.

Age (In Years)	Frequency
18-20	2
21-23	5
24-26	13
27-29	25
30-32	22
33-35	8
36-38	6
39-41	3
42-44	5
45-47	3
48-50	3
51-53	2
54-56	2
57-59	1

The cumulative frequency and the cumulative relative frequency distributions are as follows.

Ages (Years)	Frequency	Cumulative Frequency	Cumulative Relative Frequency
18-20	2	2 =2	2/100 = 0.02
21-23	5	2+5= 7	7/100 = 0.07
24-26	13	7+13 = 20	20/100 = 0.20
27-29	25	20+25= 45	45/100 = 0.45
30-32	22	45+22= 67	67/100 = 0.67
33-35	8	67+8= 75	75/100 = 0.75

36-38	6	75+6 = 81	81/100 = 0.81
39-41	3	81+3 =84	84/100 = 0.84
42-44	5	84+5 = 89	89/100 = 0.89
45-47	3	89+3= 92	92/100 = 0.92
48-50	3	92+3= 95	95/100 = 0.95
51-53	2	95+2= 97	97/100 = 0.97
54-56	2	97+2= 99	99/100 = 0.99
57-59	1	99+1= 100	100/100 = 1.00

4.0 Conclusion

In this unit you have learned how to arrange data in ordered form. You have also learned how to summarize data, both grouped and ungrouped data, into frequency distribution, relative and cumulative frequency distributions. You have also learned some basic concepts that will aid the analysis of grouped data latter in subsequent units.

What you have learned in this unit will facilitate further presentation and analysis of the data in the subsequent units.

5.0 Summary

In this unit, we have summarized data using frequency distribution, relative and cumulative frequency distributions. Some concepts that will aid the graphical presentation of the unit that will follow this unit were also defined.

6.0 Tutor Marked Assignment

For the table presented below prepare

(i) The frequency distribution for the class interval 11-20, 21-30, 31-40.....
(ii) Cummulative Frequency distribution.
(iii) Relative Frequency.
(v) Cummulative Relative Frequency Distribution

Ages of 100 Employees

58	37	21	28	27	24	27	39	38	30
60	20	30	50	44	33	23	26	31	32
18	23	41	32	27	42	40	29	28	34
56	19	28	29	23	47	29	24	31	34
30	19	22	53	49	26	16	38	42	36
41	32	33	20	31	36	32	32	31	32
21	24	55	21	24	34	37	29	33	32
32	49	38	48	33	43	26	38	30	28
24	46	15	43	43	23	23	28	34	28
41	23	25	19	51	23	36	31	35	31

7.0 References, Further Readings and other Resources

Daniel W. W. and Terrel J.C (1979) Business Statistics: Basic Concepts and Methodology 2°d ed. Houghton Mifflin Co. Boston

Harper W. M. (1982) Statistics 4`t` ed. MacDonald and Evans.

Levin R.I. (1990) Statistics for Management 4`' ed. Prentice - Hall of India Private Ltd.

UNIT 3:
GRAPHICAL PRESENTATION OF DATA

CONTENTS

1.0 Introduction

You learned in Unit 3 the summarizing of data using frequency distribution, relative and cumulative relative frequency distributions. In this unit, you will

learn how to represent these data graphically. Graphical representation has a way of showing the salient features of data without having to interpret the column of numbers. Some other data that are not summarized into frequency distributions will be represented with charts in this unit.

2.0 Objectives

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

- Plot histogram, frequency polygon and Ogive for a set of values. • Construct simple, component and multiple bar charts for a set of data.
- Draw pie chart

3.1 Histogram:

3.1.1 One of the ways of representing a frequency distribution is by means of a histogram. In constructing a histogram we plot the frequencies of the class intervals against the class boundaries [not the class limit]. The vertical axis is used for the frequencies and the horizontal axis for the class boundaries.

Example 4.1

Suppose the table below shows the distribution of 50 spectators in a secondary school sports competition. You are required to represent the data with a histogram

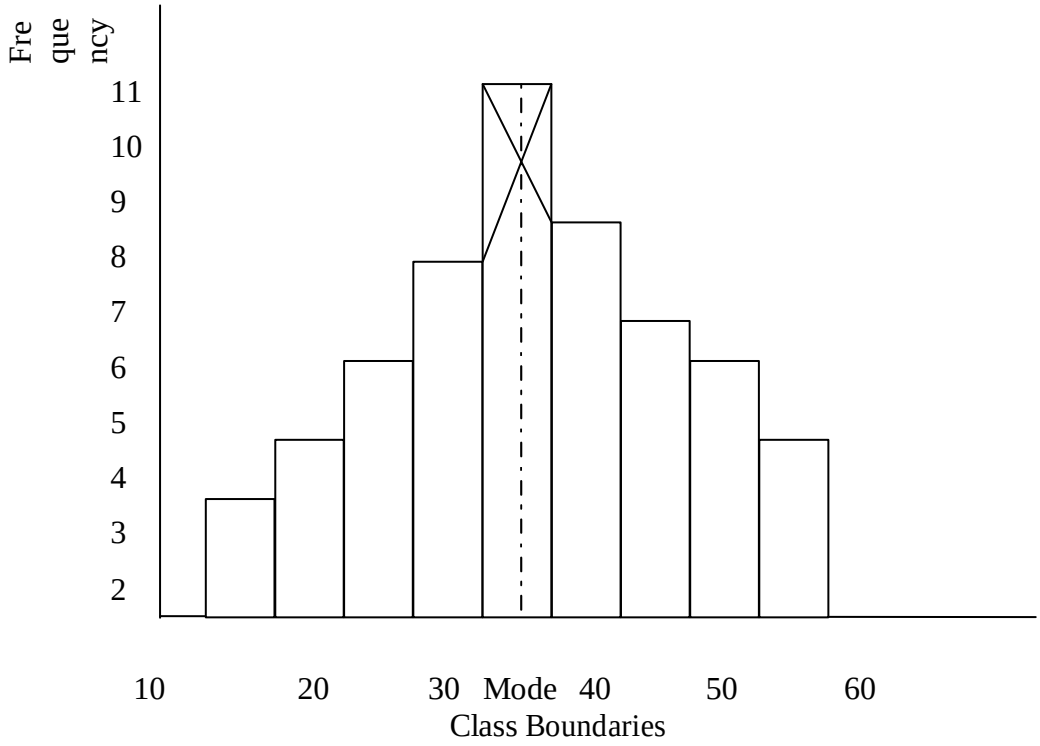
Table 4.1: Age Distribution of spectators in school sports.

Age (Years)	Frequency
10-14	2
15-19	3
20-24	5
25-29	7
30-34	11
35-39	8
40-44	6
45-49	5
50-54	3

In solving the exercise, you require a graph sheet; you also need to prepare the class boundaries for the class intervals. You will then plot the frequency against the respective class boundaries. You still need to recall how the class boundaries are computed in unit 3 we will not discuss it here.

Age (Years) Class Interval	Class Boundaries	Frequency
10-14	9.5-14.5	2
15-19	14.5-19.5	3
20-24	19.5-24.5	5
25-29	24.5-29.5	7
30-34	29.5-34.5	11
35-39	34.5-39.5	8
40-44	39.5-44.5	6
45-49	44.5-49.5	5
50-54	49.5-54.5	3

FIGURE 4.1 Histogram Showing The Data In Example 4.1



The histogram can be used to estimate the mode of the distribution. To do this you have to locate the highest cell in the histogram, join the upper class boundary of the cell with the upper boundary of the preceding cell, join the lower class boundary of the highest cell with the lower class boundary of the succeeding cell, locate the intersection, draw a vertical line from the intersection to the horizontal. The value of the vertical line on the horizontal axis is the mode. You need to see the construction on the histogram in fig. 4.1. The mode read from figure 4.1 is 32.5.

3.1.2 Frequency Polygon.

Another way of representing frequency distribution graphically is by the means of a frequency polygon.

In constructing a frequency polygon, we plot the frequency against the class marks. You learned in unit 3 that class mark of a class interval is the mean of the lower and the upper class boundaries or limits of the class interval.

Example 4.2

Construct a frequency polygon for the data in example 4.1

To construct the frequency polygon, you need to compute the class mark for the class intervals, you need to make the polygon touch the horizontal at both ends. To do this, you have to compute the class mark for an imaginary class interval at the beginning and another imaginary class interval at the end of the distribution.

If you look at table 4.1 that is of interest here, there is no class interval 5-9 at the beginning and there is no class interval 55-59 at the end of the distribution. We need to bring these intervals in and assign a frequency of 0 to each of them.

Let us now compute the class mark for the class intervals.

Class Intervals	Class Marks	Frequency
5-9	7	0
10-14	12	2
15-19	17	3
20-24	22	5
25-29	27	7
30-34	32	11
35-39	37	8
40-44	42	6
45-49	47	5
50-54	52	3
55-59	57	0

The next activity is to plot the frequency against the class marks. The frequency is on the vertical axis and the class mark on the horizontal axis. Frequency polygon must be plotted on a graph sheet

FIGURE 4.2: Frequency Polygon for Example 4.2

3.1.3 **Ogive**

This is another way of representing frequency distribution graphically. The other name for ogive is cumulative frequency distribution curve. This curve is very important in the determination of median, quartriles, percentiles, semi-interquartile range, that will be discussed in some subsequent units.

In plotting ogive for a distribution, you will do the following

(a) Compute the upper class boundaries of all the classes including that of an imaginary class at the beginning of the distribution.
b) Prepare a cumulative frequency distribution for the data
(c) Plot the cumulative frequency against the upper class boundary.

Example 4.3.

For table 4. l, construct an ogive for that distribution.

Class Interval	Frequency	Less than	Cumulative Frequency
5-9	0	9.5	0
10-14	2	14.5	2
15-19	3	19.5	5
20-24	5	24.5	10
25-29	7	29.5	17
30-34	11	34.5	28
35-39	8	39.5	36
40-44	6	44.5	42
45-49	5	49.5	47
50-54	3	59.5	50

You will realize that the class interval 5-9 was introduced. A frequency of 0 was also assigned to the class interval since the original table did not show the class. This is done so that the ogive can take its origin from the horizontal

line. From the table you will see that all the values that are less than 24.5 are contained jn class interval 5-9, 10-14, 15-19 and 20-24. The sum of the values which is equal to the cumulative frequency of the interval is 0+2+3+5 = 10.

You should know that ogive can only be plotted on a graph sheet.

FIGURE 4.3: Ogive Showing the Frequency Distribution in Table 4.1

Cummulative Frequency

Exercise 4.1

The frequency distribution of salaries (monthly) of workers in a firm is as follows

Salaries	Number of Workers
2,000-7,000	5
7,000-12,000	8
12,000-17,000	24
17,000-22,000	5
22,000-27,000	3
27,000-32,000	2
32,000-37,000	1

For the distribution, prepare

(i) a histogram and
(ii) an ogive

Using the histogram, estimate the mode of the distribution

3.2 Bar Chart

3.2.1 Simple Bar Chart

Another way of representing data graphically is by means of bar chart. A bar chart shows vertical bars with equal width to represent the values of a variable in some intervals of time. The area of the bar is proportional to the magnitude of the quantity it represents. The bars must be drawn on graph sheet and they must have equal width. There are simple components and multiple bar charts. These will be discussed in this unit.

In a simple bar chart, we use the bars to represent the value of a variable in a period of time.

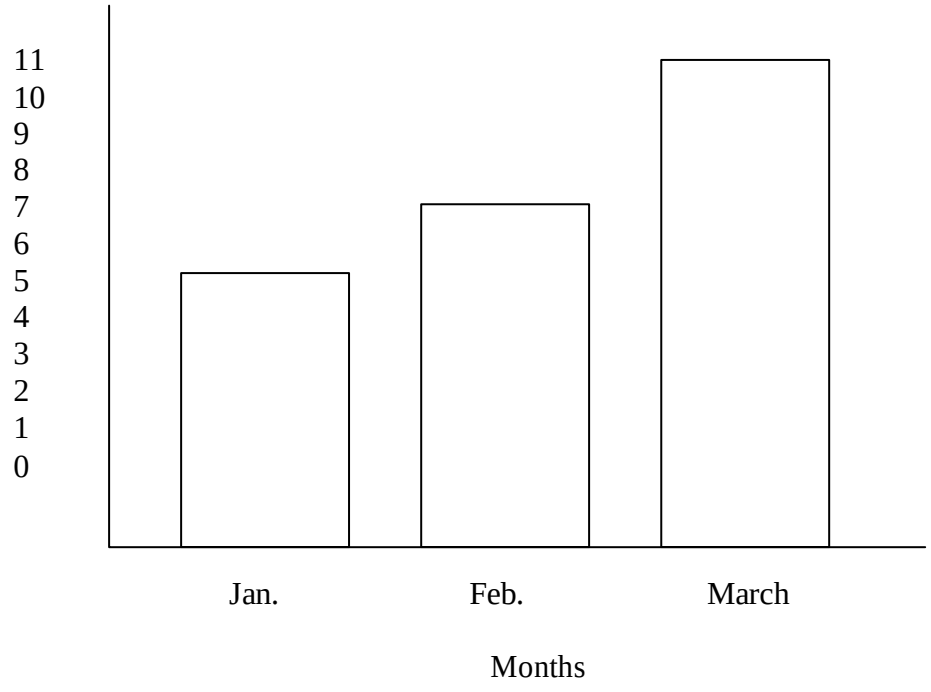
Example 4.4

Suppose the monthly sales of a firm for three consecutive months are given as follows

Months	Sales (In million of Naira)
January	5.2
February	7.4
March	10.6

You are required to represent the data on a simple bar chart.

Figure 4.4 Simple Bar Chart for Example 4.4



3.2.2 Component Bar Charts

Another bar chart that shows the total value for a time period and the values of the components that makeup the total is the component bar chart. In this case, the bar for the total value for a period is divided into the values for the components that make up the total.

Example 4.5

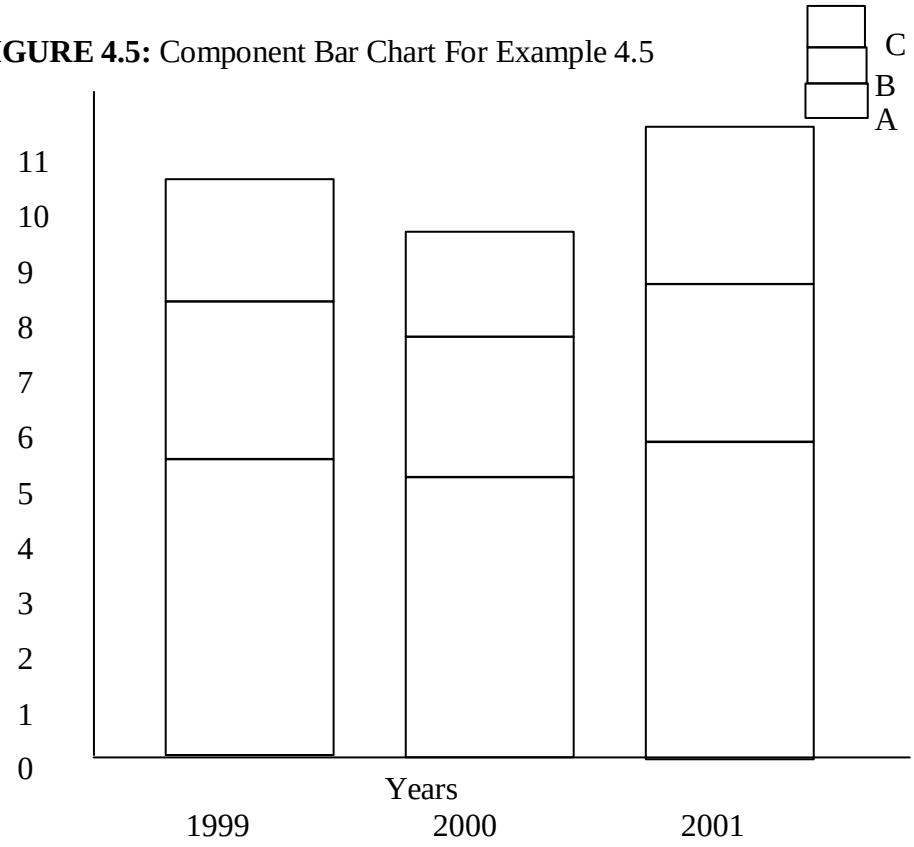
Suppose a hotel has three departments A, B, C from where sales are made and the annual records of the net profit of the departments for three consecutive years are as presented below.

Years	Net Profits in the Departments (Millions of Naira)		
	A	B	C
1999	3.2	3.4	2.8
2000	2.8	3.0	2.6
2001	4.0	3.2	3.6

You are to represent the values of the net profits with the aid of a component bar chart.

You will need to plot the values of the components A, B, and C for three years. The first year will show a bar of length 9.4cm divided into 3.2cm for A, 3.4cm for B, and 2.8cm for C. You will then repeat the exercise for the values of A, B and C in years 2000 and 2001. There must be a legend to show the shading of the component.

FIGURE 4.5: Component Bar Chart For Example 4.5



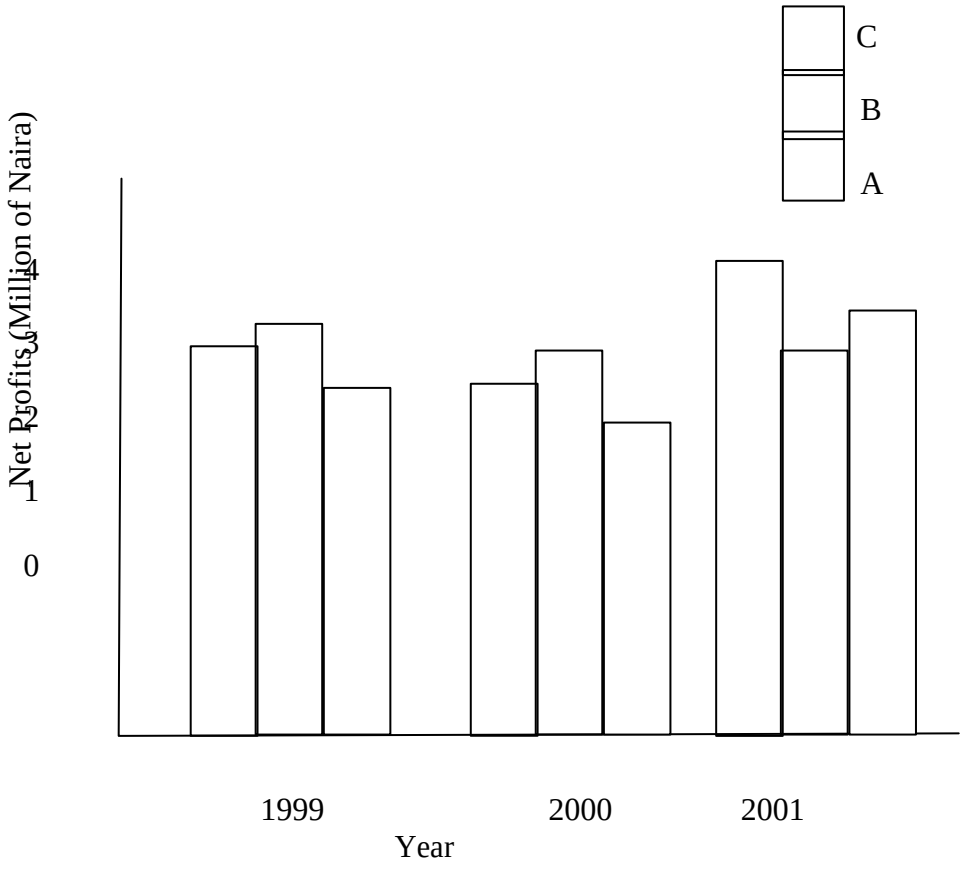
3.2.3 Multiple Bar Chart

For the multiple bar chart, each component of every year is presented by a bar whose length is corresponding to the value of the component.

Example 4.6

Using the values of net profit in example 4.5, construct a multiple bar chart for the hotel for the period three years.
In this exercise, you will draw single bar for each component of every year. The bars for a year will now look like histogram.

FIGURE 4.6: Multiple Bar Chart for Example 4.6



3.3 Pie Chart

This is another means of representing data graphically. The values of the Items represented with pie chart are proportional to the area of the sectors that represent them.

In the case of pie-chart, the sectorial angles are computed for the items based on their values and on the total values of the items.

$$\text{Sectorial angle of an item} = \frac{\text{Value of Item}}{\text{Total Value of all items}} \times 360^{\circ}$$

After obtaining the sectorial angles for the items we use a pair of compasses, a pencil, protector and ruler to draw the angles of the sector.

Example 4.7

Suppose the monthly income of a worker is allocated as follows

Items	N
Feeding	9625
Rent	4125
Education	5500
Savings	6875
Others	1375
TOTAL	27,5000

You are required to represent the data on a pie char. We will therefore compute the sectoral angles

Items	Sectoral Angles
Feeding	126"
Rent	54"
Education	72°
Savings	90"
Others	18'
TOTAL	360

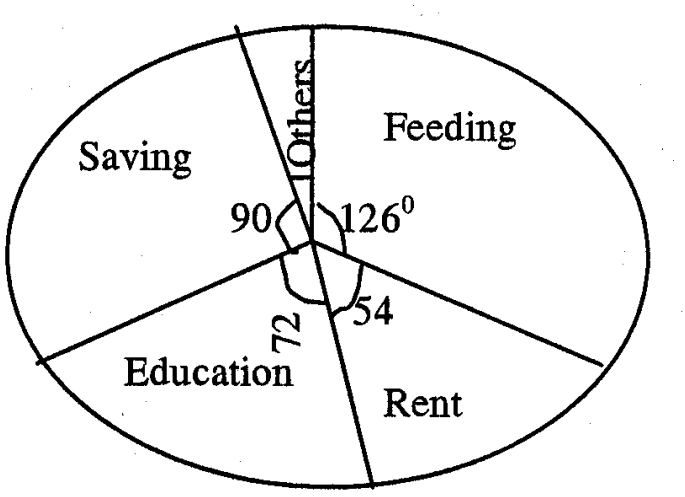
For instance the sectoral angle for feeding is calculated as

$\frac{9625}{126^0} \times 360^0 =$

27500

The total values of the items is N27, 500. This is to say that the monthly income of the worker is N27, 500. 360° is used in the calculation because the sum of angles at a point is 360°. If the sectoral angles are computed correctly, the sum of the angles must be equal to 360°.

FIGURE 4.7: Pie Chart showing Items in Example 4.7



It is possible the values of the items are given as percentages of the total values of the items. To find the sectoral angles, we only need to multiply the respective percentage with 360°

Example 4.8

Suppose the basic elements of cost of a restaurant are expressed as percentages of sales as follows.

Elements	Cost as % Sales
Food Cost	40
Labour Cost	35
Overhead Cost	15
Net Profit	10
Sales	100

Obtain the sectoral angle for elements of cost.

Sectoral Angles

Food Cost = $\frac{40}{100} \times 360^0 = 144^0$

Labour Cost = $\frac{35}{100} \times 360^0 = 126$

Net Profit = $\frac{1}{1} \times 36^0 = 36^0$

4.0 Conclusion

You learned in this unit the graphical representations that summarize the data further after the frequency distribution has been constructed. You learned that histogram is constructed by plotting frequency against the class boundaries [not class limits]; frequency polygon is constructed by plotting frequency against class mark and the ogive or cumulative frequency curve is constructed by plotting cumulative frequency against the upper class boundary.

You also learned how to represent other forms of data with bar charts and pie charts. Of all these graphs and charts only pie chart can be drawn on plan sheet of paper, all the others are to be constructed on graph sheets.

5.0 Summary

This unit taught graphic representation of frequency distributions in terms of histogram, frequency polygon and ogive. The unit also taught graphical representation of other forms of data with bar and pie charts. Subsequent units will teach the computation of measures of central tendency from ungrouped and grouped data.

6.0 Tutor Marked Assignment

6.1 The salaries of some staff in a department are given as follows: 4 clerical staff earn N1000 and more but less than N4000 each. 6 supervisors earn N4000 and more but less than N7000 each. 8 managers earn N7000 and more but less than N10, 000 each, 4 senior managers earn N10, 000 and more but less than N13, 000 each while 3 General managers earn N13, 000 and more but less than N16, 000 each. Using this data draw

(i) An Ogive (ii) An histogram, to represent the distribution

6.2 The Profit made by the three income centers of an hotel for a period of three consecutive years are as follows:

Income Centers	1998	1999	2000
A	8.0	9.0	11.5
B	5.0	5.5	6.5
C	3.0	4.5	6.0

Represent these data on:

(i) A simple bar chart
(ii) A component bar chart
(iii) A multiple bar chart

7.0 Reference and Other Resources

Ajayi, J.K. (1997) Elements of Business Statistics. Unpublished Monograph Ondo State Polytechnic, Owo

Daniel W.W and Terrel J.C (1979) Business Statistics: Basic Concepts and Methodology 2"a ed. Houghton Mifflin Co. Boston

Levin R.T (1990) Statistics for Nonageomeric 4`" ed., Prentice-Hall of India Private L.t.d. New Delhi.

UNIT 5

RESEARCH METHODS

1.0 INTRODUCTION

Ordinarily, a Research is a process comprising of many stages. Therefore careful planning is very necessary if a research is to be successful and useful.

Our first task and which we will accomplish in this unit is to examine the concept of a research. This concept is very important as it gives us an overview of research methods. The knowledge so gained, will lead us throughout the duration of this course.

2.0 OBJECTIVES

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

- explain the meaning of research
- describe the different types of research and plan.

3.0 MAIN CONTENT

3.1 Meaning of Research

Research in this context is defined from the perspective of statistical survey. It is defined as an investigation Into a particular field of enquiry. It is an organised inquiry carried out in order to provide information needed to solve a problem. Alternatively, it is defined as a systematic and objective search for and analysis of information relevant to the identification and Solution of any problem.

Research is a process comprising of many stages. Therefore careful planning is very necessary if a research is to be successful and usefuL The stages or steps involved in research process are discussed below.

3.2. STEPS IN RESEARCH PROCESS

- Clear and concise statement of the problem.
- Formulation of the research questions.
- Specification of the aims and bjecrives of the research.
- Definition of terms.
- Research design
  - determination of the target popularmon.
  - method of the sample selection.
  - dererrnination of the sample size.
  - determination of the method of obseniatson
  - desigising the tool for data collection or research instrument.
  - Pre-testing the research instrument.
  - specifying the method of analysis to be used.
- Data collection.
- Data processing.
  - editing
  - coding
  - sorting
  - tabulation
  - presentation
  - analysts
  - interpretation of results
- Report writing

(1) Clear and concise statement of the problem

The first step in every research is to first of all, specify clearly and concisely what the problems are. The statement of the problem is the motivation of the research. It should not be too vast or too narrow.

(2) Formation of research questions

Research questions are questions posed such that answers to the questions provide solution to the problem at hand.

(3) Aims and objectives of the research

In planning a research or designing a survey, the aims and objectives should first be clearly outlined. They serve as guide in the course of the survey as attempts are made to achieve the outlined or set objectives.

(4) Definition of terms

It is pertinent to define terms or words used especially where the terms are technical in nature. This enables the user of the published result understand better the message or content of the research.

(5) Research design

(a) Determination of the target population of the survey

The population of interest should be clearly determined. The nature, aims and objectives of the research determine which population should be targeted.

(b) Method of sample selection

The technique of sample selection in the survey could be said to be the backbone of the research. The better the techniques, the better the research results. Random or probability sampling techniques are usually advocated as they reduce bias to the barest minimum.

(c) Determination of the sample size

The size of the sample depends to a large extent on the size of the population. The choice of the sample size should be made in such a way that the sample would be the very representative of the entire population.

(d) Determining the method of data collection

Here we consider and select the most appropriate method of data collection suitable for our purpose.

(e) Designing the instrument for data collection.

In designing the questionnaire for the survey research, care must be taken to include all relevant questions that will solicit information on items in the aims and objectives of the research. The questionnaire must meet all requirements of a good questionnaire.

(f) Pretesting the data collection instrument

After the design of a questionnaire, it should be subjected to test in other to ascertain its workability, validity and reliability. This is done by interviewing very few respondents and analyzing their responses to see if the questionnaire is capable of achieving the result set far the research.

(g) Specifying the method of analysis

The appropriate method of analysing the data that will be generated in the courses research need be specified.

(6) Data collection

This is the data collection proper. The researcher, using the adopted method of data col. far the research, collects the data from the field or laboratory using the instrument designed purpose.

(7) Data processing

This is the whole process of editing, coding, sorting, tabulation, presentation, analysis interpretation of results. Data processing may be manual, or electronic.

Editing: This is going through the field returns to check for omissions, gaps, mis-recording etc.

Coding: This is the assigning of numerical values to nominal items in the questionna:re. For ex question on sex may have response as either male or female. For convenience, we may decide to 1 to male and 2 to female. This is called coding. Also level of education cart be represented a none, 1 for primary, 2 for junior secondary, 3 for senior secondary, 4 for polytechnics and university.

Tabulation: this is the representation or classification of the collected information in tabular form Examples are the frequency tables, contingency tables, etc.

(8) Report writing

This is writing a technical report of all the activities of the research and the interprets statistical test and manipulations. Also it entails interpretation and meaning of research results. Reporting gives the reader of the research report all the necessary information about the research the steps taken as discussed above, must be included in the report.

SELF ASSESSMENT EXERCISE 1

1. What do you understand by the term “Research”
2. What are the purposes of research?

3.2.1 TYPES OF RESEARCH

Research can be grouped in any of the following broad groups

1. Basic or fundamental or Pure Research
2. Specific applied research
3. General applied research

Basic Research

This is also known as fundamental or pure research. This research seeks to extend the boundaries of knowledge in a given area with no necessary immediate application of the result research to existing problems. It contributes to general laws and theories relating to our knowledge and understanding of the universe including behaviour of individuals, groups and organizations.

Specific Applied Research

Specific applied research s the name implies is a research conducted specifically to particular problem. it contributes to immediate decisions and actions by providing informs or understanding about a particular issue or problem. It is a research directed to needs and goals and its results expected to be useful and relevant.

General Applied Research

This attempts to use existing knowledge as an aid to the solution of some given problem or set of problems. It combines both theory (or model) building and testing with a focus en issues of general be Interest. For instant, Boolean algebra was used hundred years after it was developed, as a foundation for the logical design of electronic digital computers.

Assumptions of Scientific Research

(a) Order

Research process must be orderly. The steps or stages must be orderly arranged. For instance, the problem must first be known and understood before we set out the objectives.

(b) Determinism

Problems have causes, determinants, or antecedents that can be detected. Determinism assumes that events are related and that this relationship can be detected.

© Parsimony

This means simplicity. Simple explanation of events is required so that all probable factors can be recognised from such explanation.

(d) Empiricism

This means that the conclusions or findings of the research must be based on the data and empirical experimental results, which are testable.

(e) Generalization

This possible to generalise the research findings for the entire population. If probability sampling techniques are used in data selection, it is possible to generalise the sample result for the entire population.

(f) Specificity

Operational concepts must be specific to the research.

(g) Empirical verification

Assumption, principles and equations must be verified through the collection and manipulation of empirical data. The method of verification must also be specified.

Purpose of Research

The purpose and importance of research are outlined as follows:

(a) ft helps to increase the knowledge and understanding of problems, events or a process.

(b) Research entails problem- solving. Knowledge of research provides training n problem — solving techniques.

(c) Research provides reliable and valid information which is very important in planning and development.

(d) Research helps in the discovery of principles on which interpretation, explanation, predictions, control of behaviour can be based.

(e) It provides valid and dependable data which is useful to explain and support theories and guess.

(f) Research findings and discoveries when implemented leads to improvement.

3.3 USES OF RESEARCH TO AN ORGANISATI0N

Corporate planning

Research is used in corporate planning in order to make decisions about what goals the organisation should have in both the short and long term. Therefore research is used for;

+ Forecasting the size of future demand and trends for the organisation's products.

+ Identifying markets to be served

• Assessing the strength and weakness of the organisation both absolutely and relatively to competitors.

• Measuring dissatisfaction and needs.

+ Identifying industry/ market structure and composition

+ Identifying the strength of competitors.

• Market share and profitability analysis.

+ Highlighting significant problems.

+ Stimulating research for new product or exploitation of existing products and markets by planned policies.

+ Evaluating corporate identity and image.

+ Selecting companies for acquisition or divestment

Market planning

Research is used in market planning to keep the organisation in touch with its markets by;

+ Identifying, measuring and describing key market segments' behaviour and attitudes.

+ Assessing relative profitability of markets.

+ Analysing business potential of new market area

+ Identifying and evaluating markets for products,

4 Measuring consumer preferences.

4 ldentifying changes in competitive activity.

+ Forecasting sales.

Product planning and packaging

Research is used in making and adapting products to fulfill consumer wants more accurately and profitably. This is done by:

+ Generating and screening new product ideas and modifications.

+ Testing concepts.

+ Product testing and re- testing for acceptance and improvement.

+ Testing formulation and presentation preferences packaging teats

4 product name tests.

+ Test marketing.

+ Comparative testing against competitive product.

+ Product elimination or product line simplification

Promotional planning

Research is used for the selection and testing the effectiveness of persuasive communications

in the following areas;—

(i) Sales force planning

This includes:

+ Determination of sales areas

+ Testing alternative testing techniques.

• Setting sales target

• Evaluation of sales performance

(ii) Advertising planning

Advertising planning entails:

+ Message design and content

+ Pre- testing advertising

4. Post- testing advertising, e.g. awareness, recall, attitude shifts, brand switching effects

4 Advertising weight-of- expenditure tests

+ Media selection, media scheduling, media research

+ Advertising effectiveness

(iii) Distribution planning

Research is used to optimize the effectiveness of distribution policy. Research helps in the folov.ing ways:

Channel selection

Distribution cost analysis

Assessing distribution achievement

Determining penetration levels

For stock check

Formulation of inventory policy

(iv) Price planning

Research can be used in determining the price of a commodity. It helps in the determination of wholesale: and retailer margins and optimal cost of product and distribution. The determination of these enhances price selection for the product.

SELF ASSESSMENT EXERCISE 2

Discuss three sources of new project ideas.

4.0 CONCLUSION

This unit has treated the concept of research methods which is a stepping stone into our study of research project. Now that we have built the necessary background, we shall be familiar to research methods in project analysis.

5.0 SUMMARY

In this unit we have discussed the concept of the research methods. We have seen that it starts from the project research idea stage, goes to the identification stages, steps and types of research. From the research meaning stage, it moves to the planning stage and finally to the conclusion stage.

6.0 TUTOR-MARKED ASSIGNMENT

1. What are the steps in research process ? Briefly discuss each of them.
2. What are the main sources of survey error?
3. what are the assumptions of scientific research?
4. Of what use is research to a constructing firm?

7.0 REFERENCES/FURTHER READINGS

JUDE I.E, MICAN & EDIITH Statistics& Quantitative Methods for Construction & Business Managers.

MODULE TWO: CENTRAL TENDENCY

UNIT 1
MEASURE OF CENTRAL TENDENCY I THE ARITHMETIC MEAN

Table of Contents

1.0 Introduction

In unit 4 above, you learned how to construct graphs from frequency distribution and charts from raw data. The graphs and the charts gave us the trend and the patterns in the data. As beautiful and useful as these graphs and charts are, they do not give us any precise understanding of the reality of the information contained in the data.

We need single measures that can describe certain characteristics of the data and that give us a more precise understanding of the information the graphs and charts can give. This will help us make quicker and better decisions without the need to consult our original observation.

One of such measures is the arithmetic mean In this unit you will learn how to compute the mean from ungrouped and grouped data and how to compute weighted mean. You will also learn the characteristics of the arithmetic mean.

2.0 Objective

At the end of this unit, you should be able to compute the mean of the grouped data with or without a frequency distribution, and

- Compute the mean of grouped data
- List the properties, advantages and disadvantages of the arithmetic mean
- Compute weighted mean for some data

3.1 The Arithmetic Mean

3.1.1 The Arithmetic Mean of Ungrouped Data

We use average many times to mean ~he arithmetic mean. We compute arithmetic mean for both ungrouped and grouped data. We also compute arithmetic mean, which we henceforth in this unit call mean, for both the sample and the population from where the sample is drawn. Mean computed for the sample is called a statistic and it is donated by $\overline{x}$. The mean computed for the population is called a parameter and is denoted by μ You should note in this course that any measure computed for the sample is called statistic and any measure computed for the population is called parameter.

The mean of ungrouped data is the summation of the values in the set of data divided by the number of values in the set of data.

For a sample the number of values is denoted by n, that is the sample size, and for the population the population size is given as N.

Mean of ungrouped data (if a sample) is $\bar{x}$

$$\text{Where } \bar{x} = \frac{\sum_{i=1}^{n} \times i}{n}$$

$\sum$ = Summation

$\times i$ = Different values of a variable

n = Sample size

Example 5.1

Compute the mean for the following:

Scores: 7, 5, 8, 10, 11, 6, 3 ,

4, 10, 9,13, 2.

Mean= $\bar{x} = \frac{7+5+8+10+11+6+3+4+10+9+13+2}{12}$

$= \frac{88}{12} = 7.33$

You will realize that there is no frequency distribution for this example. Suppose there is a frequency distribution for the values of a variable, then how do we calculate the mean?

This is simple. If we are computing the mean for a sample, that is x, the mean

$\bar{x} = \frac{\sum fixi \text{ or } \sum fixi}{\overline{\sum fi} \quad \overline{n}}$

fi = different frequencies of values in the set of data

xi = different values in the set of data

$\sum fi$ or n is equal to the number of values in the set of data.

Example 5.2

The table below gives the frequency distribution of the mark scored by 20 students in a test conducted in Statistics for Management

Scores	Frequency
11	2
12	3
13	5
14	7
15	2
16	1

You are required to calculate the mean of the scores

To solve the problem, you should

(i) Multiply each score by its respective frequency to obtain $fixi$

(ii) Add up the products of each score and its respective frequency to obtain $\sum fixi$

(iii) Add up all the frequencies to obtain $\sum fi$ or n.

(iv) Divide $\sum fixi$ by $\sum fi$ to obtain the mean.

Scores (xi)	Frequency (Fi)	Fixi

11 12 12 14 15 16	2 3 5 7 2 1	22 36 65 98 30 16
Total	20	276

$$\text{Mean} = \bar{x} = \sum_{i=1}^{n} \frac{\sum fixi}{\sum fi} = \frac{276}{20} = 13.35$$

There can be another method of computing the mean apart from using

$x = \frac{\sum fixi}{\sum fi}$

we use the assumed mean, A, then the mean $x = A + \frac{\sum fidi}{\sum fi}$

where A = Assumed mean

di = xi-A for all the values of xi

Example 5.3: For the example 5.2, let us compute the mean again using an assumed mean of 13.

We will then have the table below for the solution of the problem.

Scores (Xi)	fi	di = Xi – A	fidi
11	2	11-13 =-2	-4
12	3	12-13 = -1	-3
13	5	13-13 = 0	0
14	7	14-13 = 1	7
15	2	15-13 = 2	4
16	1	16-13 = 3	3
Total	20		7

Where Assumed mean A = 13

Mean = $x = 13 + \frac{7}{20} = 13 + 0.35 = 13.35$

You will see that with the method of assumed mean, we should obtain the same value of mean we had before.

3.1.2 The Arithmetic Mean of Grouped Data

You have learned how to compute the arithmetic mean of ungrouped data.

When the values are many in a set of data, there is the need to group them into class intervals. You learned about this in unit 3. We need to take some time to compute the mean of grouped data.

Example 5.4

The earning per share (in kobo) of some firms is presented below with the frequency distribution.

Earning per share (in kobo)	Frequency
65-69	3
70-74	4
75-79	11
80-84	15
85-89	9
90-94	5
95-99	3

You are required to calculate the mean of the distribution.

To solve this question, we need to compute the class mark for the class intervals. The class mark becomes the Xi we will use in the computation. Immediately this is done, the whole distribution is reduced to the form of an ungrouped data with frequency distribution. You should recall that class mark is the mean of the upper and lower class boundaries (or Limits) of a class interval.

Class Mark I	Frequency	∑XIfI
Xi	fi	Fixi
67	3	201
72	4	288
77	11	847
82	15	1230
87	9	783
92	5	460
97	_3	291
TOTAL	50	4100

Mean x = 4100 = 82

50

The average earning per share of the firms is 82 kobo.
From the computation, you will still realize that the mean of a grouped data is given as

$$\frac{\sum fi \times i}{\sum fi}$$

x = the xi here are the various values of the class marks of the class

intervals.

The assumed mean method can also be used here.

Example 5.5: Using the assumed mean method, compute the mean of the distribution in example 5.4

We will still make use of the frequency distribution of the class mark

Class Mark xi 67	Frequency fi 3	A=77 di=xi-77 -10	 Fidi -30
72	4	-5	-20
77	11	0	0
82	15	5	75
87	9	10	90
92	5	15	75
97	3	20	60
TOTAL	50		250

We assume a mean of 77

$$\bar{x} = A + \frac{\Sigma fidi}{\Sigma fi}$$

$$\bar{x} = 77 = \frac{250}{50} = 82$$

We still obtain the same mean as before.

Example 5.1

For the distribution below, compute the mean, using the two methods we have used in this unit.

Class Intervals	Frequency
0-9	4
10-19	6
20-29	18
30-39	11
40-49	6
50-59	5

3.2 The Weighted Mean

Some times we are interested in showing the relative importance of some items. When we do, we attach weights, apart from the real values of the items. For instance in the Open University, some courses in a programme are core courses while the others are auxiliary courses. Core courses tend to have more credit unit than the others.

When we attach weighs to items we can only compute the weighted mean.

Example 5.6

The weighted mean is given as: $\bar{x}_w =$

Where xi = values of the items
Wi = relative weights of the item
$\bar{X}_w$ = Weighted mean

$$\frac{\sum wi \times i}{\sum wi}$$

The percentage scores of a student in some courses and the credit units of the courses are given below:

Courses	% Score	Credit Unit
HEM 101	52	3.0
HEM 102	64	3.0
HEM 103	71	3.0
HEM 104	45	3.0
GNS 101	82	2.0
GNS 126	55	1.0

Calculate the weighted percentage mean for the scores. xi = Percentage Scores

Xi = Percentage Scores
wi = Credit Units

Xi	wi	Wixi
52	3.0	156
64	3.0	192
71	2.0	142
45	3.0	135
82	2.0	164
55	1.0	55
	14	884

The weighted percentage mean = $\frac{884}{14}$ = 60.29

Exercise 5.2

The scores of a student in a university in the courses he took in the first semester of the year and the units of the courses are presented as follows:

Courses	% Scores	Unit
CSC 101	75	3.0
CSC 102	70	3.0
CSC103	42	2.0

CSC104	45	3.0
GNS 101	52	2.0
GNS 128	63	1.0

The University is interested in computing weighted grade point average. The academic standard of the University shows the following scores, grades and grade points

Score	Grades	Grade Point
Above 75	A	4.0
70-74	AB	3.5
65-69	B	3.25
60-64	B C	3.0
55-59	C	2.75
50-54	CD	2.5
45-49	D E	2.25
4044	F	2.0
below 40		0.0

With this information, compute the weighted grade point average for the student to two decimal places.

3.3 Advantages And Disadvantages Of Arithmetic Mean

3.3.1 Advantages of Arithmetic Mean.

The arithmetic mean has the following advantages

(i) Mean is the best known of all the averages
(ii) Mean can be used for further mathematical process. Mean is used to perform statistical procedures such as estimation and hypothesis testing.
(iii) Mean is unique, unlike mode (this will be discussed later], because a set of data has one and only one mean.

3.3.2 The disadvantages of mean

Arithmetic mean has the following disadvantages.

(i) Since all the values in a set of data are used to compute the mean, the mean can be influenced by extreme values

For instance the mean of 3, 4, 5, 6 and 7 is $\frac{3+4+5+6+7}{5} = 5$

There is no extreme value here. Suppose we have 3, 4, 5, 6, 7, 19, we then have an extreme value, 19. The mean becomes

$$\frac{3+4+5+6+7+19}{5} = \frac{44}{5} = 7.33$$

The extreme value has greatly influenced the mean.

(ii) A mean may result into an impossible value where the data is discrete. For instance, If the number of female students in five programmes at the National Open University are 35, 38, 42, 53, and 66, the mean value of the female students will be

$$\frac{35+38+42+53+66}{5} = \frac{234}{5} = 46.8 \text{ Students}$$

This is an impossible value.

(iii) We are unable to compute mean for data in which there are open-ended classes either at the beginning of the distribution or at the end of the distribution. It will be difficult to know the class mark of the open-ended class.

4.0 Conclusion

This unit teaches the computation of the arithmetic mean from grouped and ungrouped data. It also teaches the computation of the weighted mean. Some advantages and disadvantages of the arithmetic means were given in the unit. Subsequent unit will teach other measures of central tendency.

5.0 Summary

You learned in this unit the computation of arithmetic mean and weighted mean. The advantages and disadvantages of mean were also learned. Future

units will discuss other measures of central tendency.

6.0 Tutor Marked Assignment

From the frequency distribution table shown below

(i) Compute the mean using the two methods in this unit
(ii) Compare your result and comment

Class	Frequency
8.0-8.9	5
9.0-9.9	7
10.0-10.9	10
11.0-11.9	13
12.0-12.9	15
13.0-13.9	5
14.0-14.9	3
15.0-15.9	2

7.0 References, Further Reading and other Resources

Daniel W.W and Terrel J.C (1979) Business Statistics: Basic Concepts and Methodology 2nd ed. Houghton Mifflin Co. Boston

Harper W.M (1982) Statistics. 4th ed. Macdonald and Evans

Levin R.1 (1990)Statistics for Management 4^{th} ed. Prentice - Hall of India Private L.t.d

UNIT 2:

MEASURE OF CENTRAL TENDENCY2, GEOMETRIC MEAN AND HARMONIC MEAN

CONTENTS

1.0 Introduction

In unit 5 you learned low to compute the Arithmetic mean, weighted mean and the advantages and disadvantages of arithmetic mean. You also learned that the arithmetic mean is the most commonly used of all the means. Many times we are dealing with some quantities that change over a period of time. Examples of such cases are growth rates in the population, the growth factors of interest values.

In all those situations, the computation of the arithmetic mean maybe inappropriate. The appropriate means in those situations are geometric and harmonic means. This unit will discuss these types of mean. Examples will be worked. Exercises will also be provided to show your understanding of the unit.

2.0 Objective

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

Compute harmonic mean for a set of values
Compute harmonic mean for a set of values
List the uses of geometric and harmonic means.

3.1 GEOMETRIC MEAN

3.1.1 Computation Of Geometric Mean

For values $x_1, x_2, x_3, \ldots x_n$ the geometric mean is the nth root of the product of the values. The geometric mean is denoted as GM therefore

$$GM = \sqrt[n]{x_1, X_2, x_3, \ldots x_n}$$

Where GM is geometric mean $x_1, X_2, x_3, \ldots x_n$ are values of the variable of interest, while n represents the sample size.

(i) Multiply the values all together to get the product
(ii) Take the nth root of the product

Example 6.1

What is the geometric mean of 3, 5, 6, and 7?

$GM \sqrt[4]{3x5x6x7x}$

$= \sqrt[4]{630}$

$= 5.01$

When the values are many this may be difficult to compute. We can also compute the geometric mean as follows;

The formula for geometric mean is:

$GM = \sqrt[n]{x1.x2.x3\ldots xn}$

We can take the logarithm of both sides and obtain log GM = long GM

$= \sqrt[n]{x1.x2.x3\ldots xn}$

Under this situation we have

$$\text{Log GM} = \frac{f2\ \log x1\ + f2\ \log x2\ \ldots fn\ \log xn}{n}$$

If the value of frequency distribution, the geometric mean will be

$$\text{Log GM} = \frac{f2\ \log x1\ + f2\ \log x2\ \ldots fn\ \log xn}{\Sigma f1}$$

After getting the values of the logarithm and we have done the addition of all the logarithmic values we will take the anti-log of the value to obtain the geometric mean.

Example 6.2

Using logarithm, compute the geometric mean of 3, 5, 6 and 7

$GM = \sqrt[4]{3x5x6x7}$

$$\text{Log GM} = \frac{\log 3+\log5+\log6+\log7}{4}$$

Log3 = 0.4771

Log5 = 0.6990

Log6 = 0.7782

Log7 = 0.8451

Total = 2.7994

$$\text{Log GM} = \frac{2.7994}{4} = 6998$$

Anti-log 0.6994 = 5.01

Suppose there is a frequency distribution for the value. We can still compute the geometric mean.

Example 6.3

Compute the geometric mean for the distribution

X_i	Fi
5	2
6	3
7	6
8	4
9	3
10	2

In computing the geometric mean we have GM =

$\sqrt[20]{5^2 \times 6^3 \times 7^3 \times 8^4 \times 9^3 \times 10^2}$ we can use logarithm to solve the problem. Since we know that:

Log GM = $\frac{2 \log 5 + 3\log 6 + 61\log 7 + 41\log 8 + 31\log 9 + 21\log 10}{20}$

x_i	fi	Log x_i	$f_i \log x_i$
5	2	0.6990	1.398
6	3	0.7982	2.3346
7	6	0.8451	5.0706
8	4	0.9031	3.6124
9	3	0.9542	2.8626
10	2	1.0	2.000
			17.2782

Log GM = $\frac{17.2782}{20}$

= .86391

anti-log of 0.86391

= 7.31 The geometric mean is 7.31

Let us discuss some practical examples of geometric mean.

Example 6.4

Suppose the rates of growth in the sales of a product over a period of 5years are 8%, 10%, 12%, 18% and 24%. If the sales in year 0 is N100, 000,000 what is the sales at the end of the fifth year?

We may wish to find the mean rate of growth as

$\frac{8 + 10 + 15 + 18 + 24}{5} = 14.4\%$

Having got this we can get the sales at the end of the fifth year to be equal to 100,000,000 x 1.134 x 1.144 x 1.144 x 1.144 x 1.144 =N195,943,166.6.

You will see that we use the factor of 1.144 when the rate of growth is 14.4%. This is correct because we will need to add 100% and 14.4% to get the new value for a year

$$100\%+14.4\% = \frac{100}{100} + \frac{14.4}{100} = 1.144$$

The factor obtained here using arithmetic mean may not be appropriate. The appropriate factor should be obtained using geometric mean.

The appropriate growth rate is:

$$GM = \sqrt[5]{1.08 \times 1.1 \times 1.12 \times 1.18 \times 1.24} = 1.1425$$

The true growth rate is therefore 14.25% and not 14.4%. That we got initially. The sales at the end of the fifth year will then be equal to 100,000,000 x 1.1425 = N194,687,539.2

3.1.2 Uses of Geometric Mean

The geometric mean is used mainly in connection with index numbers. It is also used for arranging ratios.

Exercise 6.1

The table below shows the percentage increase in the net worth of a business over a period of 6 years.

Year	1994	1995	1996	1997	1998	1999
C/O increase	4%	9.5%	8.0%	5%	65%	75%

What is the average increase in the net worth over the period of 6 years?

3.2 HARMONIC MEAN

3.2.1 Computation Of Harmonic Mean.

For values $x_1, x_2, x_3 \ldots x_n$

The harmonic mean is given as

$$\frac{n}{\sum \frac{1}{x_1}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots \frac{1}{x_n}}$$

For this we do the following:

(1) We add up the reciprocals of the values
(2) Divide the sum into the number of items

Example 6.5

Compute the harmonic mean of 5, 6 and 7

$$\text{Harmonic mean} = \frac{3}{\frac{1}{5} + \frac{1}{6} + \frac{1}{7}}$$

$$= \frac{3}{0.2 + .0167 + 0.143} = \frac{3}{0.5099} = 5.88$$

The harmonic mean can also be used to obtain the average of different speeds

Example 6.6

A motorist moves from point p to Q, a distance of X kilometers with a speed of 100km/hour and returns to point P with a speed of 80km/hour. What is the average speed?

$$\text{The average speed} = \frac{2}{\frac{1}{100} + \frac{1}{80}} = \frac{3}{.00225} = 88.89 \text{ km/hour}$$

3.2.2 Uses Of Harmonic Mean

The harmonic mean is used to average ratios, speeds etc. It is used mostly in engineering.

Exercise 6.2

Compute the harmonic mean of the following values
5, 7, 9, 11, 13, and 15

4.0 Conclusion

In this unit, you have learned how to compute geometric and harmonic means. Geometric mean has been given for values $x_1, x_2, x_3, \ldots x_n$ as $\sqrt[n]{x_1.x_2.x_3...x_n}$ and as

Log GM = log XI + log x2 + log x3... + log xn n

The harmonic mean has been given for values xl, X2, x3, ... Xn as n n

$$\frac{n}{\sum \frac{1}{x1}} = \frac{n}{\frac{1}{x1} + \frac{1}{x2} + \frac{1}{x3} + \ldots \frac{1}{xn}}$$

Some of the uses of geometric and harmonic means have been given in this unit. Progress exercises are also given. The future units will still teach some other measure of central tendency.

5.0 Summary

This unit has shown the computation of both the geometric and harmonic means. It has also given some uses of these means

6.0 Tutor Marked Assignment

6.1 The rates of inflation in four consecutive years in a country were 6%, 8% 16% and 30%

What was the average rate of inflation per year?

6.2 Find the harmonic mean of 8, 9, 12, 15

7.0 REFERENCES AND OTHER RESOURCES

Hannagan T.J (1982) Mastering Statistics. The Macmillan Press Ltd.

Harper W.M (1982) Statistics 4'h Ed. Macdonald and Evans Handbook Series

UNIT 3:
MEASURE OF CENTRAL TENDENCY3,
MEDIAN AND MODE

CONTENTS

1.0 Introduction.

In the last two units, you learned about some measures of central tendency, their computation, their advantages and disadvantages. In this unit you will still learn about more measures of central tendency, their computation, advantages and disadvantages. We are often very interested in what is happening to the majority of the population and we show concern for this,.

In this unit two important measures, median and mode, concerned with what happens to the center of the set of values will be worked and exercises will be provided to test your understanding of the content of the unit

2.0 Objectives

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

- Compute the median for both the ungrouped and grouped data.
- Estimate median from the cumulative frequency curve.
- Compute the mode for both the ungrouped and group data.
- Estimate the mode from a histogram.
- List the advantages and the disadvantages of median and mode.

3.1 MEDIAN

3.1.1 Computation Of Median From Ungrouped And Grouped Data.

The median is a measure that shows the most central item in a set. It is a single value computed from the set of data that measures the central item in the data. If median is the most central item in the set of data, half of the values in the set must lie below the median, and the other half above the median

To calculate the median, there is a need to have a mathematical definition of the median

Median = the $\left(\frac{n+1}{2}\right)^{th}$ item in the set of data. Where n is the number of values in the set of data.

For an ungrouped data with odd number of values, the item in the middle constitutes the median.

Example 7.1

What is the median of 17, 12, 13, 15, 18? These values are not arranged in any order. To find the median, we need to order the arrangement of the values. as 12, 13, 15, 17, 18. The value in the middle is 15 and that is the median.

Using our formula for median

$$\text{Median} = \text{the } \left(\frac{n+1}{2}\right)^{th} \text{ item}$$

-

5 + 1
(

th

) 2 item

$= 3^{rd}$ item

The third item is 15

Suppose we have an even number of values such as 8, 4, 3, 5,9,2,11,7., what will be the median?. We need to order the arrangement of the values as 2,3,4,5,7,8,9,11

There are 8 values. The median is therefore the $\left(\frac{n+1}{2}\right)^{th}$ item

$$\left(\frac{8+1}{2}\right)^{th} \text{ item}$$

$$= 4.5 \text{ item}$$

The 4.5^{th} item an only be between 5 and 7. Therefore the median will be equal to

$\frac{5+7}{2})^{th} = 6$

Suppose we have ungrouped data with frequency distribution, how do we compute the median?

We still compute the median the same way we did the two previous examples

Example 7.2

For the distribution below, compute the median

Scores	Frequency
2	3
3	4
4	6
5	7

6	5
7	3
8	2

The sum of the frequencies here is 30. There are people examined. The

median should now be the $\left(\frac{30+1}{2}\right)^{th}$ item
$= 15.5^{th}$

The median (the 15.5^{th} value) is $\frac{5+5}{2} = 5$

To obtain the median for grouped data, we still do the same thing we did previously.
However, we can decide to adopt the use of formula for the median.
Some authors use the formula shown below

$$\text{Median} = L1 + \frac{J}{fi} \times w$$

Where L1 = Lower class boundary of the median class
fi = frequency of class that contain the median
w= class width of class containing the median
J = $\frac{n}{2}$ minus sum of all the frequencies up to, but not including the medium class.
n = sum of all the frequencies in the set of values $\sum fi$

Some other authors uses the formula below for median

$$\text{Sample median} = \left(\frac{\frac{30+1}{2} - (f+1)}{fm} \right) w + Lm$$

Where m = total number of items in the distribution
f = sum of all the class frequencies up to, but not including, the median class
fm = frequency of median class
w = class width
Lm = lower limit of median class interval for the purpose of this course, we will continue to use

$$\text{Median} = L1 + \frac{J}{fl} \times w$$

Example 7.3

For the distribution below, compute the median

Class interval	Frequency
0-499	229
500-999	306
1000-1499	398
1500-1999	298
2000-2499	127
2500-2999	66
3000-more	26

The sum of the frequency here is 1450 the median is the $\left(\frac{1450 + 1}{2}\right)$ th item

The median class is then 1000 -1499
Since the 725.5^{th} item lies in the class. From the first two classes we obtain (229 + 306) items = 535 items.
In the third class containing 398 items, we can get 190.5 items to make up for the 725.5 items we are looking for.
Since the median class is 1000 -1499, the lower class boundary = 1000,the fi = frequency of the median class = 398.
The class boundaries for the median class interval are 1000 - 1500. The class width is equal to 500

$J = \frac{1450}{2} - (229 + 306) = 190$

$\text{Median} = 1000 + \frac{190}{398} \times 500 = 1238.69$

3.1.2 Estimation Of Median Using Ogive

We can also estimate the median by plotting ogive for the set of data.

Example 7.4

Plot an ogive for the table below and estimate the median from the ogive

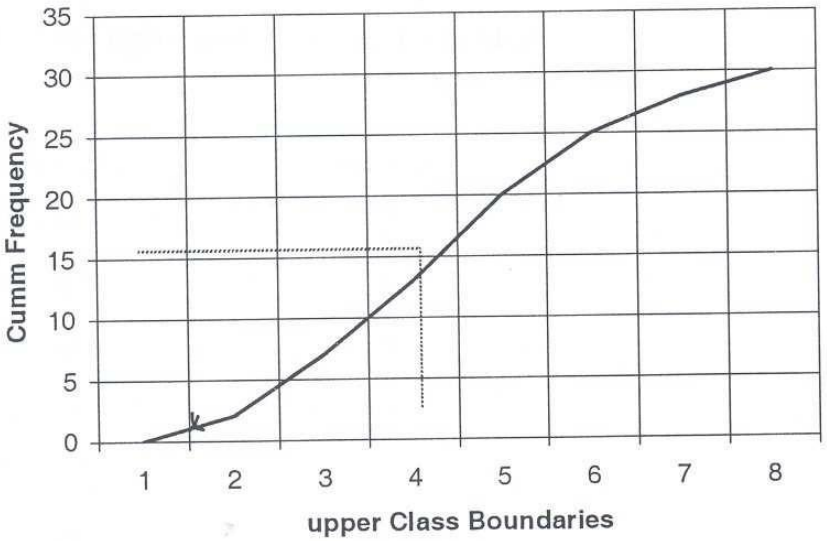
Class I	fi
1-1.99	2
2-2.99	5
3-3.99	6
4.499	7
5.599	5
6-6.99	3
7.799	2

We will need to obtain the cumulative frequency distribution to plot the ogive.

Upper Class	Cumulative frequency
1.00	0
2.00 1.00	2
3.00	7
4.00	13
5.00	20
6.00	25
7.00	28
8.00	30

To obtain the ogive, we plot the cumulative frequency against the upper class boundary as we discussed previously

FIGURE 7.1 Ogive for Example 7.4



The median is taken as $\frac{n}{2}^{th}$ the median is taken as 15^{th} value here.

Locate the 15.0 value along the vertical axis and draw it to meet the curve. At the point it meets the curve, draw a vertical line to meet the horizontal at the point the line meets the horizontal axis is the median. The estimate of the median here is 4.5 (from ogive)

3.13 Advantages and Disadvantages of median

The median cannot be influenced by extreme values as in the case of the means since not all the values are involved in the calculation of the median.

(ii) The median can be calculated from incomplete records and for open ended classes if the median does not fall into the open ended class.

(iii) It is not difficult to recognize the median

Median also has a number of disadvantages

(1) Median cannot be subjected to further statistical processing as it is with mean.

(2) To compute median we need to order the values. This can consume time.

Exercise 7.1

For the table below,

i. Compute the median
ii. Draw the ogive and estimate the median

Class Interval	Frequency
21-30	2
31-40	6
41-50	9
51-60	9
61-70	11
71-80	6
81-90	4
91-100	3

3.2 Mode

3.2.1 Computation of Mode from Ungrouped and Grouped Data

Mode is another measure of central tendency. It is simply defined as the value with the highest frequency in a set of values. Unlike mean and median, the mode can have more than one value in a set of data.

Example 7.5
From the distribution below, find the mode

X1 (ages in years I	Frequency
5	5
6	7
7	8
8	11

10	3
11	2
12	1

The mode is 8 because it has the highest frequency of 11. Suppose we replace the frequency of those in age of 10 years by 11, the distribution will have two modes, that is 8 and 10 respectively.

We can also calculate the mode from grouped data. To do this, we will use the formula.

Mode = Ll + $\frac{dixw}{di + d2}$ where Li = Lower class boundary of the modal class

d_1=frequency of the modal class minus frequency of the class just before the modal class

d_2 = frequency of the modal class minus frequency of the class just above the modal class

W = Class width

Example 7.6 For the distribution below computer the mode.

Class I	Frequency
22-26.9	5
27-31.9	8
32-36.9	14
37-41.9	23
42-46.9	19
47-51.9	7
52-56.9	4

The modal class in this question is 37-41.9.

L_1= lower class boundary = 37

W = class width = 42 -37 = 5
fi = frequency of the modal class = 23
fo = frequency of class before the modal class = 14

f2 = frequency of class after the modal class = 19

d1=fi-fo=23-14=9

d2=f1-f2=23-19=4

$$\text{mode} = 37 + \frac{9x5}{9+4} = Li + \frac{d1xW}{dl + d2}$$

= 40.46
= 40.5

3.2.2 Estimation Of Mode Using Histogram

By plotting the histogram, as was discussed in unit 4, we can also estimate the mode of the distribution.

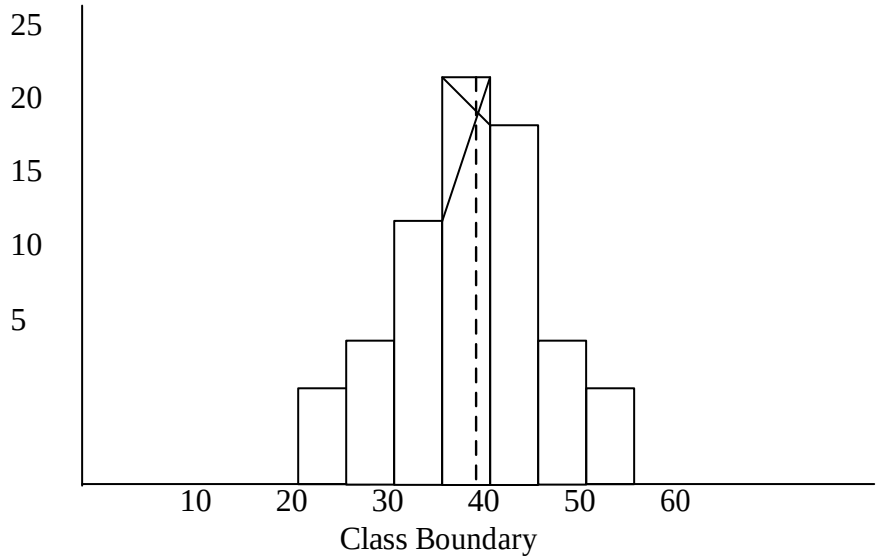
Example 7.7

Using the table in example 7.6, plot the histogram and estimate the mode.

Class Boundaries	Frequency
22-27	5
27-32	8
32-37	14
37-42	23
42-47	19
47-52	7
52-57	4

For histogram, you will recall that you still plot the frequency against the class boundaries

FIGURE 7.2: Histogram for Example 7.7



The mode obtained from the histogram is 40.0. This is a good estimate of the computed mode.

3. 2.3 Advantages And Disadvantages Of Mode

Modes has a number of advantages:

i. It can be used as central location, just as in the case of median for both qualitative date

ii. Since its computation does not involve all the values in a set of data, it can be computed from an incomplete record and it is not influenced by extreme values as it is the case with mean

Mode also has a number of disadvantages:

i It is not as commonly used, as much as means are, to measure central tendency

ii There may not be mode in a distribution if every value has the same frequency as others.

iii. There can be many modes in a distribution resulting into a problem of interpretation.

Exercise 7.2

For the table in exercise 7.1,

i. Compute the mode
ii. Draw the histogram and estimate the mode

4.0 Conclusion

In this unit, you learned that median of an ungrouped data is the median should now be the $\left(\frac{n+1}{2}\right)^{th}$ item in the distribution.

You also learned that the median of a grouped data is given by $Li + \frac{JxW}{ft}$

All the notations were properly defined in the unit. The median was also estimated from the ogive drawn from a frequency distribution.
Mode is defined as the value with the highest frequency for ungrouped data.
For grouped data, mode is defined as equal to $L1 + \frac{dixw}{d1 + d2}$

All notifications are carefully defined in the unit.
The unit gives both the advantages and disadvantages of median and mode.

5.0 Summary

The unit teaches median and mode and their computation from ungrouped and grouped data. It also shows the estimation of measures graphically.
Subsequent units will teach measure of dispersion.

6.0 Tutor Marked Assignment

From the table below:

i. Compute the median and mode
ii. Plot a histogram and estimate the mode
iii. Plot an ogive and estimate the median
iv. Compare the values in (ii) and (iii) and common.

Class intervals	Frequency
10-19	5
20-29	8
30-39	15
40-49	11
50-59	7
60-69	4

7.0 References and other Resources

Daniel W. W. Terrel J. C (1979) Business Statistics. Basic Concepts and Methodology 2n^{d} ed., Houghton Co., Boston

Levin R.1 (1990) Statistics for Management 4th ed., Prentices Hall of India Private Ltd., New Delhi.

UNIT 4

MEASURES OF DISPERSION

Contents

1.0 Introduction

In unit 5, you learned how to compute the means and indeed you did compute the means. The means of the distribution computed do not show whether the values of the distribution are clustered closely together or they spread. It is even not unusual to see two sets of values with same mean but different degrees of spread.

You realize that the per capital income (mean income/person) is important to the economist as well as the distribution of income. The per capital income may be high and majority of the populace is still very poor. While per-capital income shows the mean of incomes, the distribution of the income shows the spread.

The emphasis in this unit is on computation of variation between the values in a distribution and the mean of the distribution. This leads us to the computation of the following measures of distribution; namely, range, semi-inter quartile range, mean deviation, standard deviation and co-efficient of variation.

2.0 Objectives

At the end of this unit, you will be able, after going through the worked, examples and the exercise, to Compute the range for the distribution' Compute the semi -inter quartile range for a distribution Compute mean deviation for a distribution
Compute standard deviation for a distribution Compute the co-efficient of variation

3.1.1 Range

The range of a distribution is the difference between the highest and the lowest values in a distribution.

Example 9.1

What is the range of the following values?
2,3,5,6,8,9,11,22,34,
Range =34-2=32. The range shows how far apart the lowest and the highest value in the distribution.
It can be influenced by extreme values. It is however the simplest measure of dispersion.

3.1.2 Mean deviation

The interest here is on how far away are the values in a distribution from the mean of the distribution. The mean deviation is computed by taking the sum of the absolute value of the deviations of the values in the distribution from the mean of the distribution and dividing the sum by number of value in the distribution.

$$\text{Mean Deviation} = \frac{\sum |X_i - \bar{x}|}{n}$$

Example 9.2

Find the mean deviation of 4,6,8,9,11,12, and 13.

$$\text{The mean} = \bar{x} = \frac{4+6+8+9+11+1}{9} = 9$$

X_i	$X_i - \bar{x}$	$\sum [X_i - \bar{x}]$
4	-5	
6	-3	3
8	-1	1
9	0	0
11	2	2
12	3	3
13	4	4
		18

$|X_i - \bar{x}|$ is the absolute value of $X_i - \bar{x}$ and is equal to the positive value of the result obtained.

For example $|5-9| = 4$

$$\text{Mean deviation} = \frac{18}{7} = 2.57$$

If a group data constitutes the distribution then the

$$\text{Mean deviation} = \frac{\sum f_1 |X_i - \bar{x}|}{\sum f_i}$$

Where x_1 here is the class mark of individual class interval

Example 9.3

Compute the mean deviation of the distribution below

Class intervals	Frequency
10-14	3
15-19	5
20-24	7
25-29	3
30-34	2

The solution is as follows

x_i(class marks)	*fi*	*fixi*	$\lvert X_i - \bar{x} \rvert$	$(fi \lvert X_i - \bar{x} \rvert)$
12	3	36	9	27
17	5	85	4	20
22	7	154	1	7
27	3	81	6	18
32	2	64	11	22
Total	20	420		94

Mean = $\frac{420}{20} = 21$

Mean deviation = $\frac{\sum f_1 |X_i - \bar{x}|}{\sum f_i} = \frac{94}{21} = 4.48$

The mean deviation is a better measure of dispersion than the ranges because calculation use all the values in the distribution. It also shows how far, on the average, each of the values, lies away from the mean.

3.1.3 Quartile Deviation

In unit 8, we computed values of quartile from distribution. You may need to go back to this computation.

We have interquartile range and quartile deviation. Quartile deviation can also be called semi- interquartile range.

While interquartile range measures how far we need to go from the median on either sides before we can include one half the values of the set of data, quartile deviation or semi-interquartile range measures the average range of one-fourth of the data. Therefore,

Interquartile Range = Q3 - Q1
Quartile Deviation = $\frac{Q3-Q1}{2}$
Where Q3 = upper quartile
Q1= lower quartile

Example 9.4

The annual electricity bills in a household are given below and arranged in ascending order. The unit is in Naira.
250,540, 470, 590, 620, 650, 680, 690, 730, 750, 770, 810.
Calculate the Interquartile range and quartile deviation of the distributions.

The problem can be solved as follows:

The lower quartile the value of the
$1/4 \times 12^{th}$ item = 3rd
item. Lower = ₦570
The upper quartile is the value of the $3/4 \times 12^{th}$ item = 9"
item. Interquartile range = 730 - 570 = 160
Quartile deviation = $\frac{730 - 570}{2} = 80$

3.1.3 Standard Deviation

A measure of dispersion that enables us determine, with high degree of accuracy, where the values of a frequency distribution are located in relation to the mean, is standard deviation. We can compute it for both the sample and the population from where the sample is taken.

When it is from the samples, it is a statistic denoted by S. When it is from the population, it is a parameter denoted by ∂

The variance of a distribution is the square of the standard deviation of the distribution.

Variance (when it is sample) = 5^2

Variance (population) = ∂^2

The variance of the sample s2 is computed from this formula.
For a group of data:

$$S^2 = \frac{\Sigma f_1(X_i - x)^2}{n-1}$$

$$S^2 = \frac{n\Sigma f_1^2 \; X_1 \; (X_i - x_i)^2}{n(n-1)}$$

Where
s^2= variance of the sample
n = total number of values or sample size f_i = respective frequency of xl

x_i= class mark of the class interval for the population, the standard deviation

$$\partial^2 = \Sigma f_1(X_i - x)^2$$

Where ∂^2 = population variance
N = population size
x_i = respective values of elements in the population
μ = population mean.

Example 9.4

From the table below, compute the sample standard deviation.

Intervals	3-7	8-12	13-17	18-22	23-27	28-23
Frequency	6	8	16	10	7	3

Solution is as follows

Intervals	x_i= Class Mark	F_i	$F_i x_i$	$F_i x_i^2$
3-7	5	6	30	180
8-12	10	8	80	640
13-17	15	16	240	3840
18-22	20	10	200	4000
23-27	25	7	175	4375
28-23	30	3	90	2700
		50	815	15735

You should note that class mark (x,) is the mean of the upper and the lower class boundaries or limits.

For this example, we shall compute the variance first from

$$S^2 = \frac{n\Sigma f_i \overset{2}{X}1 (\Sigma f \times_i)^2}{n(n-1)}$$

Where $n = \Sigma f_i$

$$S^2 = \frac{50 \times 15735 - 815^2}{50(50-1)}$$

$$= \frac{78675 - 664225}{2450} = 50.01$$

The standard deviation $S = \sqrt{50.01} = 7.072$

3.1.4 Coefficient of Variation.

There may be the need to compare two distributions in terms of their variability. The standard deviation of the distribution alone can not be used to achieve comparison without knowing the means of the distribution. Again the unit of values used in the distribution may not be the same. One unit may be in kilograms another unit may be in litre. This makes comparison of the variability not easy since it is not easy to compare two things with different units.

We need to solve this problem by obtaining the coefficient of variation

$$= \frac{\partial}{\bar{x}} \times 100\%$$

Exercise 9.5

Suppose the distribution of samples A and B have $\bar{x}_B = 5.2, \bar{x}_B = 6.5$ respectively.

A has a standard deviation of 1.5 and B 2.5. Which of the distributions shows greater degree of variability?

Coefficient of variation of A $\frac{1.5}{5.2} \times 100 = 28.85$

Coefficient of variation for B = $\frac{2.5}{6.5} \times 100 = 38.50\%$

Distribution B has greater variability than A.

Exercise 9.1

For the table below compute

(a) (i) the mean (ii) mode (iii) median and (iv) standard deviation.

(b) From the measure in a above compute

(i) Coefficient of variation
(ii) Pearson Coefficient of skewness

Marks	10-19	20-29	30-39	40-49	50-59
No of Student	10	16	30	14	10

4.0 Conclusion

You learn in this unit different measures of dispersion namely, range, mean deviation, quartile deviation, interquartile range and coefficient of variations. The distance of the two values of the average from the median was the focus of this unit. A measure, coefficient of variation that determines the variability of the distribution and that is use to compare two distribution in term of their variability was discussed.

5.0 Summary

The unit reveals the need to have a better understanding on variability exhibited by distribution by computing the values of coefficient of variation, semi-interquartile range in addition to the standard deviation. The unit shows the difference in the computation of population variance and sample variance.

6.1 Tutor Marked Assignment

6.1. From the table below calculate (i) standard deviation and (ii) the coefficient of variation.

Salaries	1000-3000	3000-5000	5000-7000	7000-9000	9000-11000	11000-13000
No of workers	10	14	18	23	19	14

7.0 References And Other Resources.

Daniel W.W and Terrel J.C (1979) Business Statistics: Basic concepts and Methodology 2nd ed. Houghton Mifflin Co. Boston

Harper W.M (1982) Statistics .4th ed., Macdonald and Evans.

Levin R.1 (1990) Statistics For Management 4`h ed., Prentice-Hall of India Private Limited New Delhi.

MODULE THREE: SET THEOREM AND PROBABILITY

UNIT 1
SET THEORY

Contents

1.0 Introduction

In this unit, you will learn about some ideas and notations for set theory that will aid your understanding and calculation of probability that will constitute greater part of many of the subsequent units to study in this course. In this unit you will learn the basic concepts in the set theory and solve some problems involving the use of Venn diagrams.

2.0 Objectives

- By the end of the unit, you should be able to
- Understand set and the basic concept in set theory.
- Perform basic operations in set theory
- Use Venn Diagram to solve some problems involving set theory.

3.1 Definition and Basic Concepts in Set Theory

3.1.1 Definition

A Set is a collection of definite, distinct objects called elements or members of the set.

For the set theory, the set is written in Capital letter, Set A. After each element there must be a comma. The common bracket is not allowed.

A = { a, e, i, o, u} . This is a correct representation of Set A.
A = (a, e, i, o, u). This representation is not correct. The correct bracket is not used.
A = {a, e, i, o, u) is a Set of English vowels. Set B = {2, 3, 5, 7} is a Set of prime numbers between 0 and 10

3.1.2 Basic Concepts in Set Theory

For a good understanding of Set theory, there is a need to define some Concepts you will need in this unit.

(i) Unit Set: is a Set with only one element. Example of this are B={a}

C = {p}

(ii) Null Set: is a Set that contains no element. It is an empty set. Example is B={} or B = Ø

When you use 0 for null set, the bracket should be avoided.

(iii) Universal Set: is a Set that contains all the elements in a study or a given discussion. Examples are
S = { male, female}. This is a universal set for sexes of human beings
B = [a, b,c,d..... x, y, z].This is a universal set for the English Alphabets. Universal Set is designated as

(iv) Subset: Set A is a subset of B if and only if every element in A is contained in B. All Sets are subsets of the universal set. A null set is a subset of all sets: Look at this example

A= {a, e,i,o,u}
B= {a, b, d, e, g,i,j,m,o,p,u,r}

A is a Subset of B because all the elements in A are contained in B

(v) Equal Sets: Sets A and B are equal if and only if they contain the same elements.

(vi) Compliment of a Set: The complement of set A, designated as A^1 is a set consisting all the elements in the universal set that are not in A. For example

= {the English Alphabets} B= {a, e,i,o,u}
B compliment = B^1 = fall the consonants)

3.2 Operations In Set Theory

3.2.1 Union of Sets

The union of two sets A and B is another set consisting of all elements either in A or B or both. These are all the elements in A and B without repetition of any element. The union is designated by
Union of sets A and B is given as A B.

A B= nA+nB-nA B.

Where A B is A intersection B. This is the set of element(s) that are common to both set A and set B.

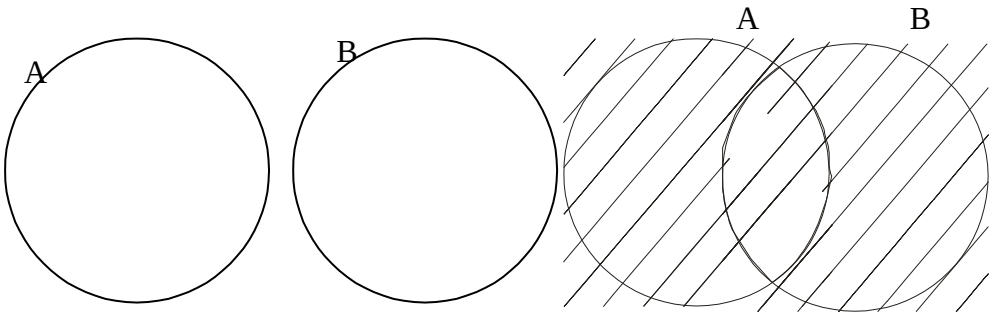
nA = number of elements in A
nB = number of elements in B
nA B = number of elements that are common to both A and B.

Example10.1

Suppose

= { Set of all integers}
A = {0, 1, 2, 3,-----}
B = {1,3,5,7,------}
C= {---6,-4,-2,0)
Find (i) A B, (ii) A C (iii) B C^1

(i) A B all elements in A and B without repetition of elements.



A B ={0,1,2,3------- }=A

B is a subset of A, therefore A B = A.

(ii) A C= f-----6,-4,-2, 0, 1, 2 3------}

(iii) For B C^1, we will need to find C^1 first. $C^1$is the compliment of C^1 C^1 is the set of all elements in the universal set that are not contained in C

C^1= {------- 5,-3,-1}

B C={------- 5,-3,-1, 1,3,5,7-------}

Example 10.2

Suppose there are 300 employees in an hotel and 120 of them have worked for the hotel for more than 5 years. What are the two sets in this hotel in terms of working years in the hotel?

A ={ 120 employees who worked for more than 5 years }

A^1= {180 employees who worked for 5 or less than 5 years in the hotel)
The two sets cannot overlap. Such sets that do not overlap are called disjoint sets.

Example 10.3

There are three factories J, K, L supplying goods to warehouses A, B, C, and D, the amount of supplies from the factories to the warehouses are shown in the table below. Obtain (i) J A, (ii) K D
(iii) C L

Warehouses / Factory	A	B	C	D	Total
J	72	16	15	50	153
K	38	18	13	22	91
L	50	32	22	43	147
Total	160	66	50	115	391

From the table

n A =160, n B=66, n C = 50, n D = 115

n J =153, n K = 91, n L=147

Also n (J B) = 16, n(K D) = 22, n(l C) = 22, n(D L) = 43

For the intersection you look at the point where the two sets intersect. Therefore

(i) J A = nJ+nA - n(J A) =153+160-72=241

(ii) K D = nK+nD - n(K D) = 91+115-22=175

(iii) C L = nC + nL = n{C L) = 50+147-22=175

3.2.2 Intersection

The intersection of Sets A and B is a set of all the elements that are common to both sets A and B.

= {Set of all integers}

A = {0, 1, 2, 3-----}

B= {1,3,5,7-----}

A B = {1, 3 ,5, 7----}

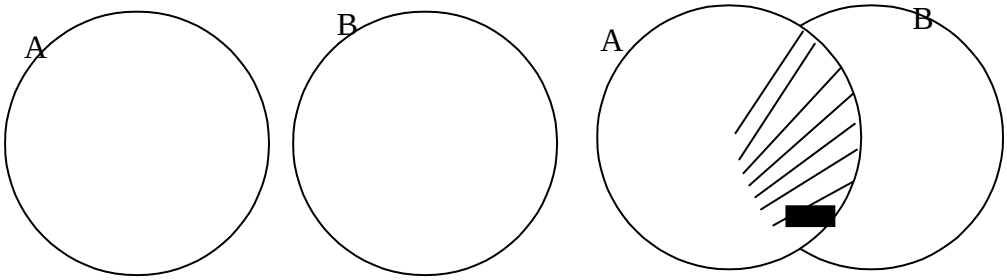
The intersection of two sets in symbolized by " "

Example 10.4

For the table in example 10.3, find. (i) A L (ii) K C (iii) C J

A L = 50
K C = 13
C J = 15

You need to know before going further that the intersection of sets A and B can be shown graphically as follows:



The shaded portion shows the intersection of sets A and B.

3.3 Venn Diagram

3.3.1 Use of Venn Diagram

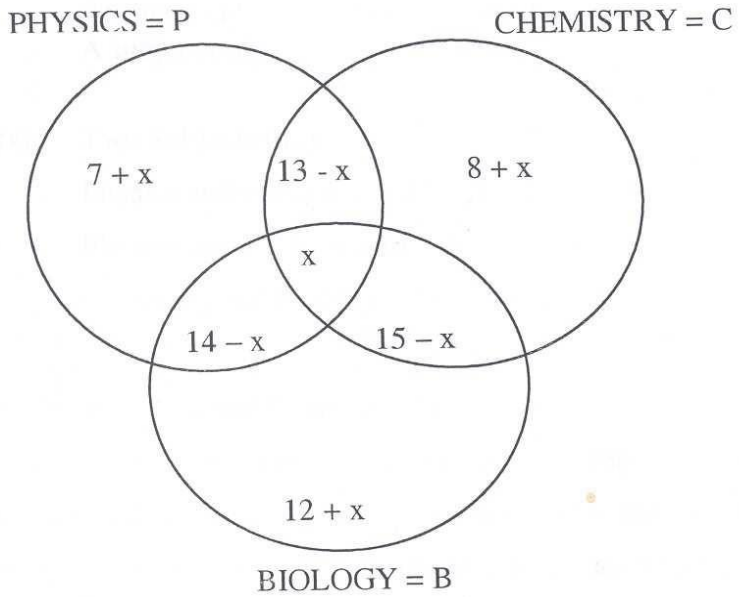
Venn Diagram is used to solve problems involving set theory.
Let us go through this aspect of set theory with an example.

Example 10.5

Suppose there are 75 students in a class with the students offering at least one of the subjects Chemistry Physics and Biology: 34 students offer Physics, 36 students offer Chemistry, and 41 students offer Biology. 13 students offer both Physics and Chemistry14 offer Physics and Biology and 15 offer

Biology and Chemistry. Represent the data on a Venn diagram and find the number that offer:

(i) All subjects
(ii) Biology only
(iii) A subject only
(iv) Two subjects only



Let the set of P C B be represented by x.
The rest part of P C = 13 – x
The rest part of P B = 14 – x

The rest part of B C=5-x

Physics only =34-[13-x+ 14-x+x] =34-13+x-14+x-x 7+x

Chemistry only = 36 - [13 - x + 15 - x + x] =8+x

Biology only = 41- [14 - x + 15 - x + x] = 12+x

Now that we have obtained the values of all the components, we will add up the values.

The sum of the values will be equal to 75.
7+x+8+x+12+x+13-x+14 x+15-x+x=75
69+x=75 x=6

(i) The number that offer the 3 subjects is 6
(ii) Biology only = 12 + x = 12 + 6 =18
(iii) A Subject

Biology only =12+x=12+6=18

Chemistry only = 8 + x-- 8 + 6 = 14

Physics only = 7 + x = 7 + 6 = 13

A subject only = 45

(iv) Two Subjects only

Physics and chemistry = 13 - X = 3 - 6 = 7

Physics and Biology = 14 - X = 14 - 6 = 8

Chemistry and Biology = 15 - X = 15 - 6 = 9

Total = 24

Student Assessment Exercise 10.1

There are 85 participants at a national Conference on Child nutrition. 35 Yoruba, 38 women and 41 nutritionists attended the conference. 13 were both Yoruba and women, 16 were both women and nutritionists and 15 were both nutritionists and Yoruba. 13 of the Yoruba were neither women nor nutritionists, using Venn diagram, Calculate:

(i) The Yoruba participants that were women and nutritionists.
(ii) The number that were neither Yoruba nor nutritionists
(iii) The number that were neither Yoruba nor women nor nutritionists

4.0 Conclusion

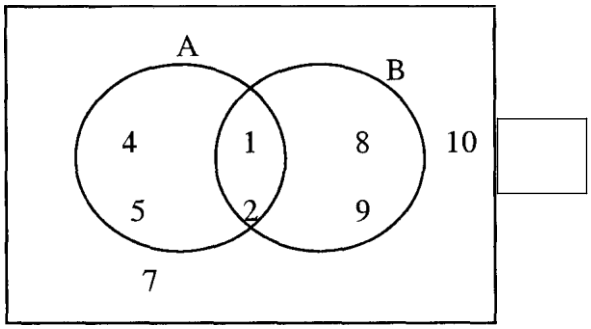
In this unit, you learned the definition of set and some basic concepts in set theory. You also learned the union and the intersection of sets. Venn diagram

is treated to solve problems in set theory. Exercises and examples given are to assist you further in the understanding of the unit.

5.0 Summary

Set has been defined as collection of definite, distinct objects called elements or members of the set. The union and intersection of sets are treated to give a good understanding of how numerical problems in set theory could be solved.

6.0 Tutor Marked Assignment



Define Sets A, B, U, A B, A B, B^1, A^l, (A B)^1

Given

A = {1, 3, 5, 9, 11, 13, 15, 17, 19}

U ={1,2,3,----------20}

C = {3,6,9,12,15,18}

D = {2, 3, 5, 7, 11, 13, 17, 19)

Find:

(i) A C (ii) A^1 (iii) (A C) n (C D)^1

7.0 References, Further Reading and Other Resources

Ajayi J.K (1997) Elements of Business Statistics, Unpublished Monograph, Ondo State Polytechnic Owo

UNIT 2
PERMUTATIONS AND COMBINATIONS

Content

1.0 Introduction

In unit 10 of this course you learned about set theory, which is an aspect of probability concepts. In this unit you will learn other aspects of probability concepts, such as permutation and combination. Many times, the focus is on computing the probability of some event or that of events. Some times, the total number of possible events is large. Under this condition, you need to have some method for counting the number of such events. The permutation and combination that this unit presents are useful techniques for counting the number of events comprising the numerator and/or the denominator of a probability.

2.0 Objectives

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

Compute factorials of values

-
- Compute permutations for objects that are different
- Compute permutation for objects that are not all different
- Compile value of combination for objects.

3.1 Factorials

3.1.1 Computation of Values of factorial

Let us define factorial. If we have a positive integer n, the product of all the whole numbers from n down to 1, is called n factorial and it is written as n! Therefore

n! = n! (n-1)(n-2)(n-3)(n-4)..... I

The values of n, since it is a positive integer are n =0, 1,2,3,4 .....∞
Infinity:.

Let us use numerical values for n

5! = 5x4x3x2x1
5! = 5x4!
5! = 5x4x3!
5! = 5x4x3x2!
4! = 4x3x2x1
4! = 4x3!
4! = 4x3x2!

Mathematically

1! = !
O! = !!

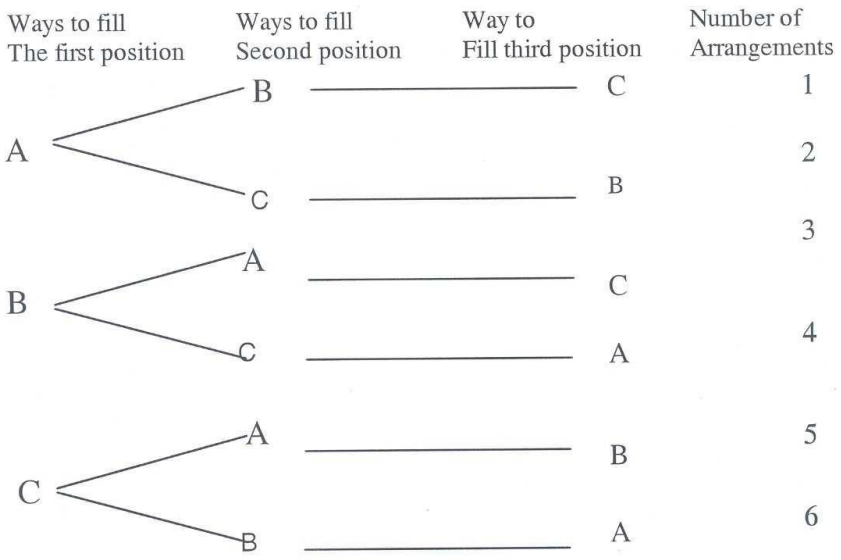
With the use of factorials, you can solve questions on the number of ways, objects or persons can be arranged in a line.

Example 2.1

Suppose there are 3 waiters on a restaurant, standing in a line against a wall, waiting for quests. How many standing arrangements are possible?
If any of the waiters can stand in any position in the line

The answer is 3! = 3x2x1=6
We can also use a graphical means to solve the problem with the aid of a tree diagram. We shall designate the positions as first, second, and third and the waiters as A, B and C, we have a tree diagram presenting the possible arrangements

FIGURE 11.1 **Tree Diagram**



We now find out the number of possible arrangements is 6.
If instead of 3 waiters, we have 4 waiters, the number of possible arrangements will become

4! = 4x3x2x1 = 24 ways

3.2 PERMUTATIONS

3.3 Computation of permutation for objects that are different

Permutation can be defined as an ordered arrangement of objects. The tree diagram Fig. 11.1 that shows possible arrangements of three objects taken three at a time gives the possible permutation of 3 objects taken 3 at a time. In some situations, as in many cases in Nigeria, there more people applying

for the number of vacancies available. For instance, we may wish to fill three positions and 15 people are applying. How then do will go about it?

There are some ways we can solve the problem. There can be the use of logical reasoning and factorial method.

We have 15 people available and 3 positions to fill. The first position can be filled in one of 15 ways, the second in one of the 14 ways and the third in one of 13 ways. The resulting possible ways will be the product of 15,14 and 13 that is 15 x 14 x 13 = 2730 ways.

You should realize that once the first position is filled, there are 14 persons remaining looking for 2 positions. When the second position is filled, there are 13 persons remaining looking for 1 position and there is one of the 13 ways to fill the third position.

The example we are considering can be solved mathematically if we make distinction between the three positions to be filled. What we are saying here is that we are having permutations of objects that are all different.

To fill three positions, when 15 persons are available, we have 15 permutation 3 symbolically we write it as

$$15\mathbf{P}3 = \frac{15!}{(15-3)!}$$

Generally, n permutation r, that is taking r items at a time from n items available is given as $\frac{nPr}{(n-r)!} = n!$

For our example

$$15\mathbf{P}3 = \frac{15!}{(15-3)!} = \frac{15!}{12!} = \frac{15xl4x13xl2!}{12!} = 2730$$

Example 2.2

Suppose there 10 different items to be stocked in the store and we have 6 bins for the storage, what are the possible arrangements?

We will now determine the permutation of 10 items taken 6 at a time

$$I0P6 = \frac{10!}{(10-6)!} = \frac{l0x9x8x7x6x5x4!}{4!} = 604{,}800 \text{ ways}$$

3.2.2. Calculation of Permutations from Objects That Are Not All Different.

You have learned the permutations of objects when all the objects are different. This is not always the case in every situation. In some instances one

or more subsets of the items cannot be distinguishable. In permutation we are required to determine how many distinguishable arrangements are possible. This situation we are discussing could be represented as follows:

$$nP\ n1,n2\ n3\ \_\_\_\_\_nk = \frac{n!}{n1\ !\ n2!\text{---}\ nk!}$$

In this case there are n objects from which the permutations is made, we have items, nr n2, ... $^{n}/_{k}$ that are distinguishable.

Example 2.3

There are 6 vegetables to be served in a restaurant, 1 of them is of type A, 2 of them of type B and 3 of them of type C. How many arrangements of the vegetables are possible? n=6

nr=1,n2=2,n3=3

$$nPn1\ n2\ n3 = \frac{n!}{nl\ !\ n2!\ n3!}$$

$$6\ \mathbf{p}\ 1,2,3 = \frac{6!}{1!\ 2!\ 3!}$$

$$= \frac{6x5x43!}{1x2xlx3!} = 60 \text{ ways}$$

Example 2.4

There are 6 positions available to be tilled by 6 persons from 3 tribes in Nigeria on equal basis. How many distinguishable arrangements are possible based on tribes?

$$6P2,2.2 = \frac{6!}{2!,\ 2!\ 2!}$$

$$= \frac{6x5x4x3x2!}{2x1x2xl\ x2!} = 90 \text{ ways}$$

Student Assessment Exercise 11.1

A sales person has 9 products to display in a trade fair but he can display only 4 at a time, how many displays can he make if the order in which he displays is important?

3.3 COMBINATIONS

3.3.1 Computation of values of combination

Combination is an arrangement of objects without regard to order. The number of combinations of n objects taken r at a time is written as n!

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}$$

$$= \frac{\text{n factorials}}{\text{r factorials ((n-r) factorials.)}}$$

For instance if we have 3 objects A, B and C and we take 2 objects at a time, the permutations will be

AB BA
AC CA
BC CB

There are 6 ways of doing the arrangement. This is $3P2 = \frac{3!}{(3-2)!} = 6$

However, in combination, where there is no regard to order

AB is the same as BA
AC is the same as CA and
BC is the same as CB hence

The number of combinations will just be 3.

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3x2!}{2!\,x!} = 3 \text{ ways}$$

Example 2.5

A food-processing firm has 8 brands of seasoning agents from which it wishes to prepare a gift package containing 5 seasoning agents. How many combinations of seasoning agents are available?

$$\binom{3}{}$$

Solution is $= {}_2 = \frac{8!}{5!(8-5!)} = \frac{8x7x6x5}{5!x3x2x1} = 56$

Student Assessment Exercise 11.2

Find the value of (i) $\binom{10}{2}$ (ü) $\binom{15}{2}$ (iii) $\binom{18}{5}$

4.0 Conclusion

This unit teaches the definitions of factorials, permutations and combinations. This unit also teaches computation of values of some factorials, permutations and combinations. It also shows the difference in the computation of permutations when the objects are all different and when the objects are not all different. Some time is devoted to this unit because of the importance of the concepts in this unit to the study of the subsequent units in this course

5.0 Summary

The unit gives the permutations of r object, from n objects as $\frac{n!}{(n-r)}$ when all the objects are different. It also gives the permutation of objects r_1, r_2, r_3 ..... r_k from n objects to be $\frac{n!}{r_1!r_2rl_3..... r_k!}$ when the objects are not all different.

The unit gives the combinations of r objects from n objects as $\frac{n!}{r!(n-r)!}$

Examples are worked in the unit and exercises are given to facilitate your understanding of the unit.

6.0 Tutor Marked Assignment

6.1 In how many different ways can MISSISSIPPI be arranged with distinction placed on letters

6.2 If we want to arrange 15 objects picking 5 at once, how many combinations are possible?

7.0 References. Further Reading and Other Resources

Ajay J.K (1997) Elements of Statistics Unpublished monograph, Ondo State Polytechnic, Owo

Daniel W.W. and Terrel J.C. (1979) Business Statistics; Basic Concepts and Methodology 2"a ed. Houghton Mifflin Co. Boston.

UNIT 3

SOME ELEMENTARY PROBABILITY CONCEPTS

Table Of Contents

1.0 Introduction

The activities of man and the firms consist of occurrence of events. While some events are certain to occur some others can never occur. Between these extreme lie most of the events of human activities. There is always a lot of uncertainty in the occurrence of these events. The unit therefore introduces you to probability that deals with concepts and measurement of uncertainty. The unit therefore provides you the opportunity to understand some of the variability and complexity of the business world.

2.0 Objectives

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

* Understand some views of probability
* Understand elementary properties of probability
* Compute probability of an event.

3.1 Different Views Of Probability

3.1.1 Definition Of Probability

Probability is defined as a value between 0 and 1 that shows the likelihood of occurrence of an event. If the value is 0, the event can never occur. For instance the probability that a Mango tree will bear an orange fruit is 0. If the probability is 1, it means the event is sure to occur. The probability that a man born of a woman will die one day is 1. Death for any man is imminent.

Many other events have the values of the probability in between 0 and 1. This raises the question of uncertainty. There are two types of probability:

a. Objective and
b. Subjective probabilities

The objective probability can be subdivided into

(i) The classical or prior probability
(ii) The relative frequency concept or posterior probability

Classical probability is defined as follows "If some event can occur in N mutually exclusive and equally likely ways and if m of these possess characteristic E, the probability of the occurrence of E is equal to $\frac{M}{N}$

(Daniel & Terrel, 1979)

That is $P(E) = \frac{M}{N}$

P(E) is probability of characteristic E.

Example 3.1

If a couple makes five trials at birth and records 3 male children. What is probability of having a success if the birth of a male child is regarded as a success.

$P(\text{success}) \frac{M}{N}$

M=3

N=5

$P(\text{success}) = \frac{3}{5} = 0.6$

The relative frequency concept is also used to define probability. For instance we have a variable x; and we can draw a frequency distribution for the variable, then the relative frequency of a value of the variable is approximately equal to the probability of occurrence of the variable.

Example 12.2

Suppose a test is conducted for twenty students in the course Statistics for Management and the results of the test are as presented in the frequency distribution table below. Calculate the probabilities of the students scoring a particular score in the distribution.

Score	Frequency
2	1
3	2
4	5
5	6
6	3
7	2
8	1

To solve the problem we need to compute the relative frequency of each of the scores.

Total number of students is 20

Relative frequency of score 2 = $\frac{1}{20}$

This is the frequency of 2 divided by the total frequency. We can now prepare the relative frequency table for the distribution as follows

Scores	Frequency	Relative frequency
2	1	0.05
3	2	0.10
4	5	0.25
5	6	0.30
6	3	0.15
7	2	0.10
8	1	0.5

From the table the probability that a student scores 6 is approximately equal to 0.15 that is $\frac{3}{20}$

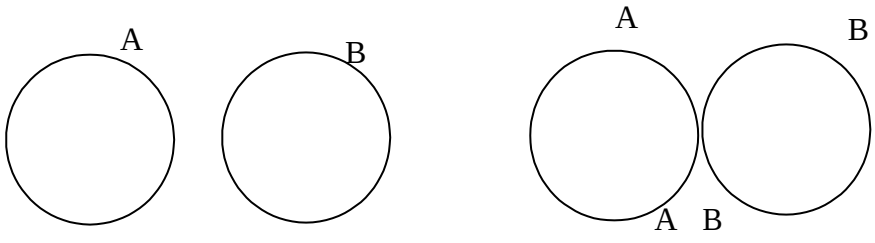
The relative frequency distribution table can also be used to find the values of many other probabilities. For instance what is the probability that a student scores more than 5. This will be the probability that a student scores 6 or 7 or 8. This will be equal to the sum of these: Probabilities = 0.15 + 0.10 + 0.05 = 0.30

3.1.2 Mutually Exclusive And Equally Likely Events

Two events are mutually exclusive when they cannot occur together. Examples of these are:

(i) Sex of a child in a birth
(ii) Values of throw when throwing a die
(iii) Head and tail in a coin
(iv) Win or loss when playing a match in which a winner must emerge.

When two events are mutually exclusive, the intersection of the events will be zero



These sets 'A and B are mutually exclusive.

Equally likely events are those events in which there is no reason to expect one event rather than the others to occur. For instance, if a coin is not biased the head and tail are equally likely. If a die is not biased, 1, 2, 3, 4, 5, 6 are equally likely to occur.

Indeed, if there is no dominant effect of heredity on the sex of a child, male and female are equally likely events.

3.2 Properties Of Probability

There are three properties of probability.

a. The first one is derived from the definition. Probability of an event is defined as a value between 0 and 1 that shows the likelihood of occurrence of the event. Therefore probability of an event cannot be negative that is less than 0. The probability can also not be more than 1. Therefore the probability of an event with characteristic E is $0 \leq P(E) \leq 1$. This is the first property of probability.

b. If events are both mutually exclusive and collectively exhaustive then the sum of the probabilities of the events is equal to 1.The events that are mutually exclusive and collectively exhaustive are as follows.

(i) Male and female as sexes of a child in a birth.
(ii) Head and tail in a throw of a coin
(iii) 1, 2, 3, 4, 5, and 6 in a throw of a die

With these examples, no other event can occur in each of the cases. The events are therefore collectively exhaustive. Since the events cannot occur simultaneously, the events are said to be mutually exclusive.

Therefore if events El, EZ, E3 ... E_n are both mutually and collectively exhaustive then

$\sum P(E_1) = P(E_1) + P(E_2) + P(E_3).-P(E_n)$
$=1$

c. The occurrence of either of two events A or B or both is given as
$P(A \cup B) = P(A) + P(B) - P(A \cap C)$
But this holds when the two events can occur simultaneously. However, if the two events are mutually exclusive then $P(A \cup B) = P(A) + P(B)$

You should note that if two events are mutually exclusive, their intersection will be zero. i.e. (A B)=0

3.3 CALCULATING THE PROBABILITY OF AN EVENT

3.3.1 Conditional And Unconditional Probabilities

It is necessary at this stage to introduce the use of the concepts introduced previously in this unit in solving of problems involving calculation of probabilities. In this section we will treat two types of probabilities - the unconditional and conditional probabilities. Let us explain these probabilities by solving a question that will bring the understanding of unconditional and conditional probabilities out properly.

Example 12.3

Suppose there are 70 Secretaries, cross-classified as by sex and marital status as follows: There are 20 male and 25 married Secretaries. 5 of the married Secretaries are male. With this information, we will prepare a table with which we can calculate the various probabilities required.

	Sex		
Marital Status	Male (S1)	Female (S2)	Total
Single (Ml) Married (M2)	15 5	30 20	45 25
Total	20	50	70

Let us now look at the unconditional probabilities. Total number of secretaries is 70. The total married secretaries are 25, total single is 45, total male is 20 and total female is 50.

With these information, we can compute the following unconditional probabilities. Probability of getting a female secretary is given by $\frac{50}{70}$

= 0.7143 similiarly

P (male) = $\frac{20}{70}$ = 0.2857

P (single) = $\frac{45}{}$ = 0.6429

P (married) = $\frac{25}{70}$ = 0.3571

All these probabilities are unconditional probabilities The conditional probability can be treated as follows; From the table, some male are single and some are married hence male and single are not mutually exclusive; so also male and married. The probability of getting a male given that he is single is given as:

P (male/single) = $\frac{P(\text{male} \quad \text{single})}{P(\text{single})}$ or

P(male/single) = $\frac{n(\text{male} \quad \text{single})}{n(\text{single})}$

Where P(male single) is probability of male intersection single and n(male n single) is the number that is common to both male and single. This probability is a conditional probability

Example 3.4

Using the table in example 3.1, calculate the following probabilities.

(i) Of getting a male given that is married
(ii) Of getting a single given that is female
(iii) Of getting married secretary given that he is a male
(iv) Of getting a female secretary given that she is single.

1. P(male/married) = $\frac{n(\text{male} \quad \text{married})}{n(\text{married})} = \frac{5}{25}$

= 0.2

Another method is P(male/married) = $\frac{P(\text{male} \quad \text{married})}{P(\text{married})}$

P(male married) is equal to the number that is common to both male and married divided by the total number of secretaries. In this example n(male married) = 5 total number of secretaries = 70 therefore P(male married) = $\frac{5}{70}$

P(married) = $\frac{25}{70}$

5

Therefore $\frac{P(\text{male wedded})}{P(\text{married})} = \frac{70}{25/70} = 0.2$

You may wish to use any of the methods, though 1 personally show preference for

$P(A/B) = \frac{n(A \cap B)}{(B)}$ for any events A and B

2 $P(\text{single/female}) = \frac{n(\text{single female})}{n(\text{female})} \frac{30}{50} = 0.6$

3 $P(\text{married/male}) = \frac{n(\text{married male})}{n(\text{male})} = 5 = \frac{0.25}{20}$

4. $P(\text{female/single}) \frac{n(\text{female single})}{n(\text{single})} \frac{30}{45} = 0.67$

Student Assessment Exercise 12.1

Suppose there are four types of appliances in the kitchen denoted by A1, A2, A3 and A4 and the colours of the appliances are of 3 types Cl, C2, C3. If the table below shows the cross-classification of the appliances by both the types and colours, Calculate the probabilities listed after the table
Colour

Types	C 1	C2	C3
Al	45	35	70
A2	60	15	25
A3	40	50	30
A4	75	25	30

From the table calculate probability that an appliance picked at random is of.

a) (i) Type Al

(ii) Type A2

(iii) Type A3

(iv) Type A4

(b) (i) Colour C_1 and Colour C_3 (c) (i) Type A_1 given colour C_3 (ii) Type A_3 given colour C_2

(iii) Colour C_2 given type A_4

4.0 Conclusion

The unit stated the properties of probability. The unit presented the calculation of the probability of an event and worked out problems involving conditional and unconditional probabilities.

5.0 Summary

You learned the three different points of view, classical, relative frequency and subjective, from which we can discuss probability. You also learned the properties of probability as:

(i) $0 \leq P(E) \leq 1$ for events E1, E2, E3 ... E_n

(ii) For mutually exclusive and collectively exhaustive events El, E2, E3 ... E_n that $\sum P(E;) = 1$

That is $P(E_1) + P(E_2) + P(E_3)$ ------ $P(E_n) = 1$

(iii) $P(E_1$ or E_2 or both) = $P(E_1) + P(E_2) - P(E_1 \quad E_2)$ but for mutually exclusive events $P(E_1 \quad E_2) = P(E_1) + P(E_2)$.

The unit also taught the conditional probability of A given that is of type B to be

$$P(A/B) = \frac{n(A \quad B)}{n(B)}$$

The future units will build on these.

6.0 Tutor Marked Assignment

The table below shows the programmes in the National Open University and the types of Employers that will employ the graduates of this programmes. It also shows the number of graduates the employer's employ.

	Types of Employer				
Programmes	Teaching	Civil Service	Manufact-uring	Merchand-ising	Others
	El	E2	E3	E4	Es
Business Admin. D1	30	20	10	14	16
Hotel & Catering D2	20	15	18	13	14
E - Banking (D3)	15	17	12	4	2
E - Library D4	15	13	25	8	9
Co-op Study D5	20	15	8	20	17
Others D6	90	5	7	11	17

From the table, compute the following probabilities

a (i) P(D4) (ii) P(D5) (iii) P(D3)

b (i) P(D5 E3) (ii) P(E2 D4) (iii) $P\left[\frac{n1}{E5}\right]$

7.0 References. Further Reading and Other Resources

Daniel W.W. and Terrel J.C. (1979) Business Statistics: Basic Concepts and Methodology 2nd ed., Houghton Mifflin Co; Boston.

UNIT 4

PROBABILITY RULE, EVENTS AND BAYES' THEOREM

Table of Contents

1.0 Introduction

In the unit 12 we discussed that probability is used to measures uncertainty in the activities of man and the firms.

There is a need to know the rules that are used in probability so that we can effectively solve problems involving probability. This unit helps us to do this. The unit will be useful in identifying some events that are important probability concepts. The unit will also discuss Bayes' Theorem that will assist the solutions to some problems.

2.0 Objectives

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

- Understand the addition and multiplication rules in probability
- Explain independent and complimentary events
- Solve problems involving joint probability and
- Apply Bayes' Theorem to solving of questions involving probability.

3.1 Probability Rules

3.1.1 Addition Rule

In unit 12, under properties of probability, you learned that the probability that event A or event B or both will occur is

P(A B) = P(A) + P(B) = P(A B).

This is true when the events are not mutually exclusive, then P(A B) = P(A) + P(B) P(A B) = P(A) + P(B) - P(A B) is known as the addition rule. When the two events are mutually exclusive P(A B) = 0

Example 13.1

Let us go back to the table we had for 3.3 in unit 12 to illustrate numerically the addition rule.

SE
X

Marital Status	Male (S 1)	Female (S2)	Total

Single (M1) Married (M2)	15 5	30 20	45 25
Total	20	50	70

The probability of picking either S_1 or M_2 or both is $P(S_1 \quad M_2) = P(s1) + P(M_2) - P(S_1 \quad M_2)$ from the table,

$P(S_1) \frac{20}{70}$, $P(M_2) \frac{25}{70}$

$(S_1 \quad M_2) = 5$

$P(S_1 \quad M_2) = \frac{5}{70}$

5 is the value at the intersection of S_1 and M_2 and 70 is the total number of people involved

$P(S_1 \quad M_2) = P(SP + P(M_2) - P(S_1 \quad M_2) = \frac{20}{70} + \frac{25}{70} - \frac{5}{70} = \frac{40}{70} = \frac{4}{7}$

3.1.2 The Multiplication Rule

This is another rule that is useful in the computation of the probability of an event.

In unit 12 under conditional probability, we discussed that the probability of event A given B is equal to $P(^{A}/_{B}) = P(A \quad B)$ for $P(B) \neq 0$

This relationship holds if P(B) is not equal to 0. If P(B) is equal to 0, the definition of a probability $0 \leq P(\Sigma) \leq 1$ will be violated. If we have

$P(^{A}/_{B}) = \frac{P(A \quad B)}{P(B)}$

Then $P(A \quad B) = P(B) \; P(^{A}/_{B})$. This is obtained by rearranging the preceding equation. P(A n B) is the probability of the joint occurrence of event A and event B.

The probability of joint occurrence of event A and event B is equal to the conditional probability of a given B that is $P(^{A}/_{B})$ times the marginal probability of B, that is P(B).

Therefore $P(A \quad B) = P(B) \; P(^{A}/_{B})$.

Example 13.2

What is the probability of joint occurrence of M1 and S2 in the table used for

example 4.1. $P(M1 \quad S2) = P(B)\ P\left[\ m1 \,/\, S2\ \right]$

$P(S2) = \left[\ \frac{50}{70}\ \right]$

$P\left[\ \frac{m1}{S2}\ \right] = \frac{(m1 \quad S2)}{n\,(S2)} = \frac{30}{50}$

$P(M1 \quad S2) = \frac{50}{70} \times \frac{30}{50} = \frac{30}{70} = -\ 0.4286$

3.2 Probability Of Events

3.2.1 Independent Events

You learned in our discussion before that P(A B)=P(B) $P(^{A}/_{B})$.

Suppose the probability of event A is the same whether event B occurs or not then events A and B will be said to be independent event therefore $P(^{A}/_{B})$ = P(A) $P(^{A}/_{B})$ = P(B)

P(A B) = P(A) P(B)

You will recall that for all cases. But if the events are independent then P(A n B) = P(A) P(B).

Example 13.3

Suppose there are 50 secretaries in the National Open University and 20 of them are single and 20 of them are male. If 8 of the male are single, what is the probability that a secretary picked at random is married given her is female.

Let us prepare the required table for this required question.

	Single (S1)	Married (S2)	Total
Male (M1) Female (M2)	8 12	12 18	20 30
Total	20	30	50

What we are interested in is to show that when events A and B are independent, the $P(^{A}/_{B}) = P(A)$

$P(\text{married}) = \frac{30}{50} = 0.6$

$P(\text{Married/ Female}) = \frac{P(\text{married} \quad \text{Female})}{P(\text{Female})}$

$P(\text{married} \quad \text{Female}) = \frac{18}{50}$

$P(\text{Female}) = \frac{30}{50}$

Therefore $P(\text{married /Female}) = \frac{\frac{18}{50}}{\frac{30}{50}} = \frac{18}{50} \times \frac{50}{30} = \frac{18}{30} = 0.6$

With this example, you will see that P(married) = P(married/ female).

It means that the fact that the married secretary is a female does not affect the probability that the secretary is married; therefore the two events are independent.

3.2.2 Complimentary Events

Two events A and B are complimentary if and only if P(B) = 1 - P(B). That is if (A) + P(B) = 1.

Examples of complimentary events are
(1) Male and female
(ii) Defective and not defective.
(iii) Failure and success.

3.2.3 Joint Probability

Having treated multiplication rule there is a need to compute joint probability of events.

Example 13.4

It is known from experience that for a particular type of seedling, 3 out of 5 them transplanted will survive. Suppose two of the seedlings are picked at random, what is the probability that:

(i) Both of them survived
(ii) The first survives and the second does not survive.
(iii) The first does not survive and the second survives
(iv) None of the them survives

Let S represents survival and f represents non - survival.

$P(S) = {}^{3}/_{5} = .6$
$P(f)=(1-0.6)=0.4$

(i) Both survive = P(S) P(S) = 0.6 x 0.6 = 0.36

(ii) First survives and second does not survive = P(S) P(f) = 0.6 x 0.4 = 0.24

(iii) First fails and second survives = P(f) P(s) = 0.4 x 0.6 = 0.24

(iv) Both do not survive = P(f)(P(f) = 0.4 x 0.4 = 0.16

Exercise 13.4

1. For the table below, compute
(1) P(S1 M1)(ii) P(M2 v S2) (iii) P(M1 n S2)

	Si	Sz
M, MZ	20 40	30 50

2. If the probability that a hunter will hit a target is $^2/_3$ and the hunter aims at the target two times, what is the probability that:

(i) He hits it the two times?
(ii) He hits it once?
(iii) He hits first and misses second?
(iv) He misses it the two times?

3.3 Bayes Theorem

The Baye theorem is stated as follows:
Given B1, B2, B3 - - - - Bn, Mutually exclusive events whose union is the universe, and let A be an arbitrary event in the Universe, such that P(A) ≠ 0 then

$$P(Bj/A) = \frac{P(A/Bj)}{P(Bj)} \overline{\Sigma(PA/Bj)\ P(Bj)}$$

Where J = 1, 2, 3------ n (Daniel & Terrel, 1979)

$$P(Bj/A) = \frac{P(A \quad Bj)}{P(A \quad BI) + P(A \quad B2) \text{----} + P(A/Bn)}$$

You should recall that (P(A B) = P(A/B) P(B)

Example 13.5

Three tutors are assigned to mark the assignments of Hotel and Catering Management Students in the National Open University. The tutor B_1 marks 45 % of the assignment, the second tutor, B_2, marks 35% of the assignment and the third tutor B_3 marks 20% of the assignment.

The course coordinator vetted the marking and it is formed that the first tutor has an error of 0.03, the second 0.05 and the third 0.04. If a script selected is found to have an error, what is the probability that, the script was marked by the first second or third tutor respectively?

We are going to apply B ayes theorem in the problem.

$P(B_1) = \frac{45}{100} = 0.45$

$P(B_2) = \frac{35}{100} = 0.35$

$P(B_3) = \frac{20}{100} = 0.20$

$P(A/B_1) = 0.03$
$P(A/B_2) = 0.05$
$P(A/B_3) = 0.04$
Let us draw a table for easy computation.

Events	P(B)	P(A/Bj)	P(A n Bj)	P(Bj/A)
B_1	0.45	0.03	0.0135	0.3462
B_2	0.35	0.05	0.0175	0.4487
B_3	0.20	0.04	0-008	0.2051
Total			0.039	1.00

You should note that

$P(A \quad Bj) = P(A/Bj)$

$P(Bj)\ P(Bj / A) = P(A$

$Bj)$
$\sum P(A \quad Bj)$

For B_1, $P(Bj/A) = \frac{00135}{0.039} = 0.3462$

B_2, $P(Bj/A) = \frac{0.0175}{0.039} = 0.4487$

B_3, $P(Bj/A) = \frac{0.008}{0.039} = 0.2051$

Exercise 13.2

Two students S_1 and S_2 are asked to record scores in an examination. Student S 1 recorded 60% of the scores and student S_2 recorded 40%. It is found that the error rates of the students S_1 and S_2 are 0.03 and 0.02 respectively. What is the probability that an error detected in the recording is committed by S_1 or S_2 respectively.

4.0 Conclusion

This unit teaches the application of addition and multiplication rules in probability. It also teaches the meaning of independent and complimentary events. Problems on joint probability are also taught. There is also the application of Bayes theorem to solving of problems in probability.

5.0 Summary

The unit reveals that

$P(A \quad B) = P(A) + P(B) - P(A \quad B)$ and that $P(A/B) = \frac{P(P(A \quad B)}{P(B)}$

for $P(B) \neq 0$

For complimentary events A and B, (P(A) = 1 - P(B)

In the application of Bayes theorem the unit gives

$$P(Bj/A) = \frac{P(A \quad Bj)}{\sum P(A/Bj)P(Bj)} \quad \text{for } j = 1, 2, \sim n) = 1, 2, \sim n.$$

$$P(Bj/A) \quad \frac{P(A \quad Bj)}{\sum P(A \quad Bj)}$$

6.0 Tutor Marked Assignment

6.1 Two hunters A and B aim at a target. The probability that A hits the target is and that B hits the target is 3/a. What is the probability that:

(i) They both hit the target.
(ii) They both miss the target
(iii) One of them hits the target
(iv) At least one of them hits the target?

6.2 In an area in Oshodi - Oke, 35% of the households use brand A, 25% of them brand B and 40% of them brand C of a seasoning agent.
The proportion of the people who learned about the product from radio advertisement is 0.03 for brand A, 0.08 for brand B and 0.05 for brand C. the household contacted reveals that the households learned about the product through radio advertisement.

What is the probability that the brand of the seasoning agent used in the household is (1) A? (ii) B? (iii) C?

7.0 References. Further Reading and Other Resources

Daniel W.W. and Terrel J.C. (1979) Business Statistics: Basic Concepts and Methodology 2"a ed., Houghton Mifflin Co, Boston.

UNIT 5
PROBABILITY DISTRIBUTION OF A DISCRETE RANDOM VARIABLE

Contents

1.0 Introduction

You learned in unit 12 of this course the basic concepts of probability theory. You also learned the method of computing the probability of an event. In unit 1 of module 1 of this course, you learned the meaning of a discrete random variable. You leaned that a discrete random variable is a variable that can only assume whole numbers.

This unit will build on the techniques of calculating probability of an event you learned before by showing how to calculate probability under a more complicated situation.

2.0 Objectives

By the end of this unit, you should be able to

- Construct a probability distribution of a discrete random variable from raw data.
- Compute the mean and variance of the probability distribution of a discrete random variable.

3.1 RELATIVE FREQUENCY DISTRIBUTION

3.1.1 Relative Frequency Distribution

In unit 12, you learned about the various views of probability. You learned that the relative frequency of a value of a variable is approximately equal to the probability of that value. This is the relative frequency definition of probability.

In that unit 12 you also learned that the probability of an event, say $X = x_{1o}$, is given as $P(x_1 = x)$ and that

$0 \leq P(x = xl) \leq l$. In the unit 12 we have made use of E and we gave that $0 \leq P(E) \leq 1$. This is the first property of probability.

Also you learned that for events x_1 x_2, x_3 ---- x_n that are both mutually exclusive and collectively exhaustive, $\sum P(x = x_1) = 1$

For this unit the probability of a value of variable x, that is equal to x will be given as P(x, = x). This should be noted throughout. If $x_1 = 5$

then we will give the probability as $P(x_1 = 5)$ Let us have a distribution of a throws of a die The events in a die are, 1,2,3 4,5,6

Example 5.1

Suppose a die is thrown 20 times and the frequency distribution of the throw is shown below.

X_1	Fl
1	2
2	2
3	4
4	6
5	4
6	2

The next thing to do is find the probability distribution of this discrete random variable. The probability of 1, given as

$$P(x_1 = 1) = \frac{\text{Frequency l}}{\text{Totalfrequency}} = \frac{2}{20} = 0.10$$

That of 2, $P(x1 = 2) = \frac{2}{20} = 0.10$

That of 4, $P(x1 = 4) = \frac{6}{20} = 0.30$

You have learned that the relative frequency of an event is approximately equal to the probability of the event.

The relative frequency distribution of the discrete random variable is given as follows:

X_1	F_1	$P(X - X_1)$
1	2	0.1
2	2	0.1
3	4	0.2
4	6	0.3
5	4	0.2
6	2	0.1
	20	1.00

With this probability distribution obtained, we can make a number of probability statements. We can even generate a number of questions from the distribution.

For instance, what is the probability that a throw will give value of 3?

The answer is $P(x = 3) = \frac{4}{20} = 0.2$.

Another one is what is the probability of getting in a throw 4 or 5. The answer is P(X = 4 or 5)

This is equal to P(x = 4) + P(x = 5) = 0.3 + 0.2 (from table) = 0.5

Another question is what is the probability that the value of the throw will be more than 3.

The answer is $P(x \geq 4) = P(x = 4) + P(x = 5) + P(x = 6) = 0.3+0.2+0.1=0.6$

3.1.2. Cumulative Relation Frequency Distribution

In view of the fact that the cumulative relative frequency distribution is important in computing the probabilities of some events, we will need to construct the cumulative relative frequency of the discrete random variable.

For a discrete random variable X the cumulative relative frequency of x_1 is given as is $P(x \leq x_1)$ for $x_1 = 0,1,2,\text{----- } n$

hence $P(x \leq 4)$ is equal to $P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4)$
For our example 5.1, the cumulative frequency is given as follow.

X_1	F_1	$P(x = X_1)$	Cumulative Relative Frequency $P(x \leq x_1)$
1	2	0.1	0.1
2	2	0.1	0.2
3	4	0.2	0.4
4	6	0.3	0.7
5	4	0.2	0.9
6	2	0.1	1.00

We can also make a number of probability statements on the cumulative relation frequency distribution.

The probability of a value that is at most 4 is given as $P(x \leq 4) = 0.7$ The probability of getting a value between 2 and 5 inclusive is given by
$P(x \leq 5) - P(x \leq 1) = 0.9\text{-}0.1 = 0.8$
Or $P(x=2)+P(x=3)+P(x=4)+P(x=5)=0.1+0.2+0.3+0.2=0.8$

You will realize that we did not say that the probability is equal to
$P(x \leq 5) - P(x \leq 2)$.

Rather we say that it is

$P(x \leq 5) - P(x \leq 1)$ because the question state that the probability of 2 is inclusive.

3.2 Measures Of Probability Distribution

3.2.1 Mean of probability distribution of a discrete Random Variable.

The mean as a concept was introduced in Unit 5. You learned that mean could be computed for both the population and the sample taken from the population. This unit treats the concept in terms of probability distribution of a discrete random variable.

The mean of the probability distribution is the expected value of the random variable that has the specified distribution.

Here we will find the expected value of X which is

$E(X) = \sum xi\ P(x = xi)$

Example 14.2

Using the table showing the frequency distribution of the values in a die thrown 20 times.

X_1	f_1	$P(x = xi)$	$X_1\ P(x \leq xi)$
1	2	0.1	0.1
2	2	0.1	0.2
3	4	0.2	0.4
4	6	0.3	0.7
5	4	0.2	0.9
6	2	0.1	1.00
		1.00	3.7

E(x)=1 x 0.1 +2 x 0.1 + 3 x 0.2 + 4 x 0.3 x 5 x 0.2 + 6 x 0.1
=0.1+0.2+0.6+1.2+1.2+1.0+0.6=3.7

You will also realize that

$X_i\ P(x = x_i) = X;\ x\ P(x = x_i)$

The E(x) obtained is the mean. If we compute it for population it is designated as μ Hence $E(x) = \mu$

Let us use the formula we used previously to compute the mean, that is

$$\mu = \frac{\sum fixi}{\sum .fi}$$

Therefore $\mu = \frac{1x2 + 2x2 + 3x4 + 4x6 + 5x4 + 6x2}{20}$

This is equal to $\frac{74}{20} = 3.7 = E(x)$

3.2.2 Variance of Probability Distribution of A Discrete Random Variable

You also learned about variance when you studied measures of dispersion in unit 9 of this course. This variance in this unit will involve that of the probability distribution of a discrete random variable. The variance is given as the mathematical expectation of
$(x - \mu)^2$ which is equal to

$$E(xl - \mu)^2 = \sum(xi - \mu)^2 P(x = xi)$$

Example 14.3

Using the table in example 5.2, compute the variance for the distribution.

Xi	fi	P(X = Xl)	(xi - μ)	(xi - μ)2	(Xi - μ)2 P(X = Xi)
1	2	0.1	-2.7	7.29	0.79
2	2	0.1	-1.7	2.89	0.29
3	4	0.2	-0.7	0.49	0.098
4	6	0.3	0.3	0.09	0.027
5	4	0.2	1.3	1.69	0.338
6	2	0.1	2.3	5.29	0.529
Total	20				2.01

The variance of the distribution is 2.01

The computation appears tedious. There is a short cut to it. It is known that E $(x - \mu)^2 = E(X)2 - [E(X)]^2$

We have computed E(x) before, hence we can easily obtain $[E(xi)]^2$

What it remains is for us to compute $E(x)^2$. This can be done as follows

Xi	Fi	P(X - Xi)	Xi	X_i^2 P(X - Xi)
1	2	0.1	1	0.1
2	2	0.1	4	0.4
3	4	0.2	9	1.8
4	6	0.3	16	4.8
5	4	0.2	25	5.0
6	2	0.1	36	3.6
Total				15.7

The $E(x)^2 = \sum x^2 P(x = xi) = 15.7$

Using $E(x - \mu)^2 = E(X)^2 - [E(X_i)]^2$

The variance, $E(x - \mu)^2 = 15.7 - 3.7^2 = 15.7 - 13.69 = 2.01$

The result is the same when we calculated the variance using

$E(x - \mu)^2 = \sum(x - \mu)^2 P(x = X_1)$

Therefore Variance = $E(x^2) - (E(x))^2 = E(x - \mu)^2$

Student Assignment Exercise 14.1

Given the table below,

X_i	F_i
0	2
1	2
2	4
3	5
4	6
5	9
6	7
7	5

8	4
9	2
10	1

(a) Prepare a probability distribution table for the data

(b) Prepare a cumulative frequency curve

(c) From the table prepared, estimate

(i) $P(3 \leq x \leq 8)$ (ii) $P(x > 5)$ (iii) $P(x \leq 4)$

(d) Compute (i) $E(x)$ (ii) $E(x2)$ (iii) $E(x - \mu)^2$

4.0 Conclusion

This unit has built more on the concepts and techniques of probability by showing the application of these concepts to the relative frequency of distribution of a discrete random variable. The unit has also giver the means of computing mean and variance of the probability distribution of a discrete random variable.

Future units will show more of the types of probability distributions both of discrete and continuous random variables.

5.0 Summary

You learned in the unit that the relative frequency of an even x; is approximately equal to its probability, $P(x = xl)$, for a discrete random variable.

The mean of the probability distribution of a discrete random variable, which is the expected value of x given as $E(x)$ is equal to $\sum xi\ P(x = xl)$. You also learned that the variance of the distribution, given as $E(xl- \mu)^2$ is equal to $\sum (x1 - \mu)^2 P(x = X1)$.

It is also discussed in the unit that $\sum(x - \mu)^2 = E(x^2) - (E(x)^2$ and this is the variance of the distribution.

6.0 Tutor Marked Assignment

The table below shows the number of industrial accidents suffered by 50 employees of a large manufacturing firm in a year.

Number of Accidents	Number of Employees
0	5
1	10
2	16
3	8
4	6
5	3
6	2

(a) (i) Construct the probability distribution for the table, that is the relative frequency distribution.

(ii) Construct the cumulative relative frequency for the distribution (P x = x1)

(b) What is the probability that a randomly selected employee will be the one who had:

(i) 4 accidents
(ii) more than 2 accidents
(iii) At least an accident
(iv) Between 2 and 5 accidents
(v) 2 or 3 accidents

(c) (i) the mean of the distribution
(ii) The variance of the distribution

(Use probability method)

Answers to Student Assessment Exercise 14.1

X_i X_1)	f_i	$P(X = x_i)$	$Px \leq X_1$	$XP(X =$	X^2_1	$X^2_1(Px=x1)$
0	2	0.04	0.04	0	0	0
1	4	0.08	0.12	0.08	1	0.08
2	5	0.10	0.22	0.20	4	0.40
3	6	0.12	0.34	0.36	9	1.08
4	9	0.18	0.52	0.72	16	2.88
5	7	0.14	0.66	0.70	25	3.50
6	5	0.10	0.76	0.60	36	3.60
7	4	0.08	0.84	0.56	49	3.92
8	4	0.08	0.92	0.64	64	5.12
9	2	0.04	0.96	0.36	81	3.24
10	1	0.02	1.00	0.20	100	2.00
				4.42		25.82

(i) $P(\leq x \leq 8) = P(x \leq 8) - P(x \leq 2) = 0.92 - 0.22 =$
0.70 (ii) $P(x > 5) = 1 - P(x <- 5) = 1\text{-}0\ .66 = 0.34$
(iii) $P(x \leq 4) = 0.52$

From the table

$E(x) = 4.42, \quad E(x_2) = 25.82$

$E(x - \mu)^2 = E(x^2) - (E(x)^2 = 25.82 - 4.42^2 = 6.2836$

7.0 References And Other Recourse

Daniel W.W and Terrel J.C (1979) Business statistics: Basic Concept and Methodology 2nd edition Houghton Mifflin Co;, Boston

Levin R.1 (1990) Statistics for Management 4th edition, Prentice -Hall of India Private Ltd. New Delhi

MODULE FOUR: CORRELATION AND REGRESSION ANALYSIS

UNIT 1: CORRELATION THEORY

CONTENTS

1.0 INTRODUCTION

Correlation can be defined as the branches of statistics that deals with mutual dependence or inter-relationship of two or more variables. If the value of two variables such that when one changes, the other changes too, then the variable are said to be correlated.

Generally, correlation implies that variation in one variable, when there is a variation in other variable.

Note that the degree of relationship which exist between two variables. The degree of relationship existing between two variables is called simple correlation. While the degree of relationship that connected three or more variables together is called Multiple correlation.

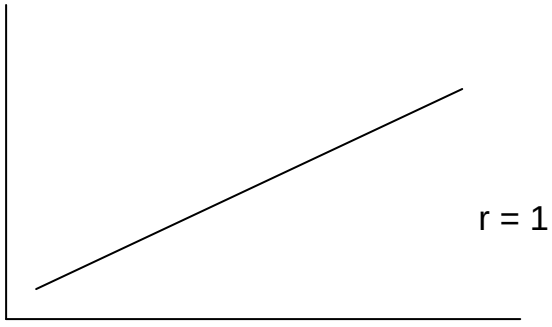
2.0 OBJECTIVES

The main objective of this unit is to enable students understand the theory behind and the application of correlation in statistics. Students are expected at the end of this unit to be able to apply correlation theory to solving day-to-day business and economic problems.

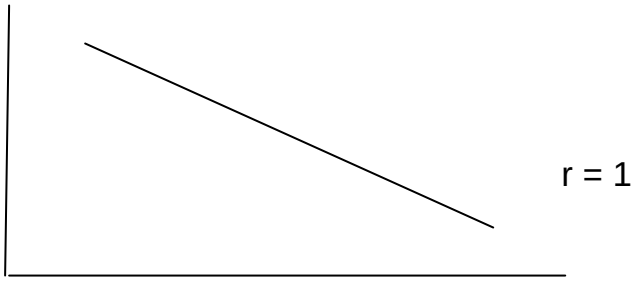
3.0 MAIN CONTENT

3.1 Perfect Positive Correlation

This can be defined as the situation where all the scatter points passes through a straight line none of the points deviated from the normal curve and positive slope.

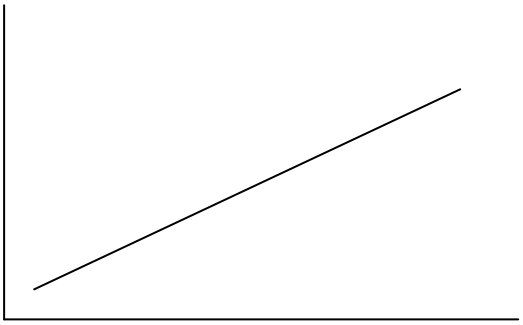


3.2 **Perfect Negative Correlation:** This indicates that all the points passes through the normal straight line and non deviated from the line. The curve shown downward slope of units.

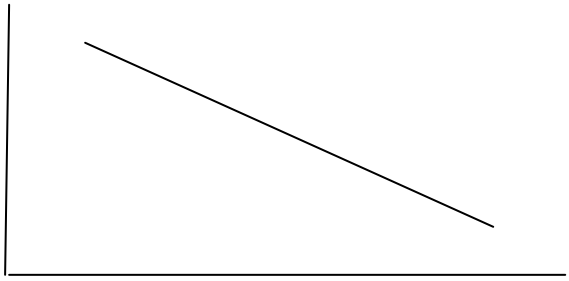


3.3 **Strong Positive Correlation:** In these case, most of the scatter points passes through the straight line, although there are few

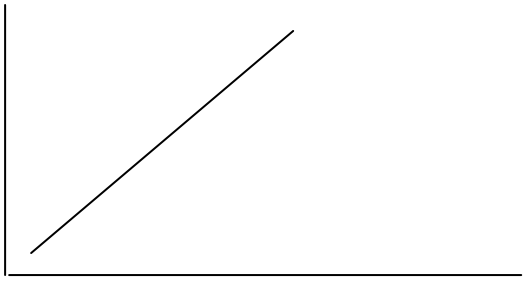
deviation from the straight line, but the deviation are very close to each other.



3.4 **Strong Negative Correlation:** In a strong negative correlation, some of the points passes through the straight line and all other scatter point are very close to the straight line, it has a negative slope which is very close to unity.



3.5 **Weak positive correlation:** In these case the points are deviated from each other so that each of the scatter points are for the depart from each other and the association is weak. The slope is positive and not close, to unity.



3.6 **Weak negative correlation:** In a weak negative correlation, there are serious deviations of scatter points and the points slope downward. It has a negative slope and not close to unity.

3.7 **No Correlation:** The scatter point at random and did not form any regular pattern for recognization by any straight line. There is no association between the variables.

4.0 Conclusion

The relationships among business variables can simply be identified using correlation coefficients. Two variables can either be positively or negatively correlated. This correlation can be linear or nonlinear depending on variable characteristics.

5.0 Summary

For a precise quantitative measurement of the degree of correlation between two variables, say X and Y, we use a parameter 〉 referred to as the correlation coefficient. The sample estimate of this parameter is referred to as r.

6.0 Tutor-Marked Assignment

1. Explain with the use of diagram different types of correlation
2. Differentiate between strong positive correlation and negative correlation.

7.0 References /Further Reading

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques

UNIT 2: PEARSON'S CORRELATION COEFFICIENT

CONTENTS

1.0 INTRODUCTION

Coefficient of correlation refers as the ratio of covariance between the related variables to the square root of the product of individual variance.

2.0 OBJECTIVE

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

- describe the computation of linear correlation coefficients
- apply the concept of correlations in business decisions.

3.0 MAIN CONTENT

Given a bivariate set of data x, y; x, y; y2 ---- x, y,

To obtain the general representations of product moment correlation coefficient as

r : $r = \frac{Sxy}{Sx\ Sx}$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 - \sum y^2}}$$

Where,

$X = x - x$

$Y = y - y$ respectively

From above equation, substitutes for x and y

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 - \sum (y - \bar{y})^2}}$$

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})(x - \bar{x}) - \sum (y - \bar{y})(y - \bar{y})}}$$

From numerator above,

$r = \sum (x - \bar{x})(y - \bar{y})$

$r = \sum (xy - xy - yx + xy)$

$r = \sum (xy - \frac{\sum y}{\cap} x - \frac{\sum xy}{\cap} + \frac{\sum x}{\cap} . \frac{\sum y}{\cap}$

$r = \sum xy - \frac{\sum y}{\cap} \sum x - \frac{\sum x}{\cap} . \sum y + \cap . \frac{\sum x}{\cap} . \frac{\sum y}{\cap}$

$= \sum xy - \frac{\sum x \sum y}{\cap}$

$= \frac{\cap \sum xy - \sum x . \sum y}{\cap}$ ----------(i)

From renominator,

$\sum (x - \bar{x})(x - \bar{x})$

$\sum (x^2 - x\,x - xx\ f\ xx)$

$\sum (x^2 - \frac{\sum x}{\cap} x - \frac{\sum y}{\cap} x + \frac{\sum x}{\cap} . \frac{\sum x}{\cap}$

$\sum x^2 - \frac{(\sum xy)^2}{\cap} - \frac{(\sum x)}{\cap} + \frac{(\sum x)^2}{\cap}$

$= \sum x^2 - (\sum X)^2$

$= \sum x^2 - (\sum x)^2$ ----------(ii)

Mathematically,

$\sum (y - y)\ (y - y)$

$\sum (y^2 - y\ y - yy\ f\ yy)$

$\sum (y^2 - \underline{\sum y}\ \ y\ - \underline{\sum y}\ x + \underline{\sum y}\ .\ \underline{\sum y}$

$\qquad\quad \cap \qquad \cap \qquad \cap \qquad \cap$

$\sum y^2 - \underline{(\sum y)^2} - \underline{(\sum y)} + \underline{(\sum y)^2}$

$\qquad\quad \cap \qquad \cap \qquad \cap$

$= \qquad \sum y^2 - \underline{(\sum y)^2}$

$\qquad\qquad\quad \cap$

$= \sum x^2 - (\sum x)^2$ ----------(ii)

Thus, equate (i) and (ii)

$= \sum xy^2 - \sum y \sum x \div \sqrt{\underline{\sum x^2 - (\sum x)^2}}\ .\ \sqrt{\underline{\sum y^2 - (\sum y)^2}}$

$\qquad\qquad\qquad\qquad\quad \cap \qquad\qquad \cap$

$\sum xy - \sum y \sum x \div \sqrt{\underline{\sum x^2 - (\sum x)^2}}\ .\ \sqrt{\underline{\sum y^2 - (\sum y)^2}}$

$\qquad\qquad\qquad\qquad\quad \cap \qquad\qquad \cap$

$= \sum xy - \sum y \sum x\ \ x\ \ \sqrt{\underline{\sum x^2 - (\sum x)^2}}\ .\ \sqrt{\underline{\sum y^2 - (\sum y)^2}}$

$\qquad\qquad\qquad\qquad\quad \cap \qquad\qquad \cap$

$r = \underline{\sum xy - \sum y \sum x}$

$\sqrt{\underline{\sum x^2 - (\sum x)^2}}\ .\ \underline{\sum y^2 - (\sum y)^2}$

Or

$$r = \frac{\sum xy}{\sum x^2 \cdot \sum y^2}$$

Remarks:

The value of r can be expressed in 3 ways of interpretation of relationship between x and y.

i. When r = +1, i.e. perfect (positive) linear relationship
ii. When r = -1 i.e. perfect (negative) linear relationship
iii. When r = 0 i.e. no relationship.

Note: The straight of relationship between x and y depends on how close r is to zero. And the coefficient of determination will be given as (r^2).

Illustration: Relationship between money spent on research and development and chemical firm's annual report profit. The information for proceeding 6 years was as recorded. Calculate product moment correlation coefficient.

Years	1994	1995	1992	1991	1990	1989	
Money (N) res and Dev.	5	11	4	5	3	2	x
Annual profit (N)	31	40	30	34	25	20	y

Data

$\mu = 6$; $\sum x = 30$; $\sum y = 180$; $\sum xy = 1000$; $\sum x^2 = 250$; $\sum y^2 = 5642$

Yrs	X	Y	xy	X^2	Y^2

1994	5	31	155	25	961
1993	11	40	440	121	1605
1992	4	30	120	16	900
1991	5	34	170	25	1156
1990	3	25	75	9	625
1989	2	20	40	4	400

$$r^1 = \frac{\sum xy - (\sum x)(\sum y)}{\sqrt{(\sum x^2 - (\sum x)^2)(\sum y^2 - (\sum y^2)}}$$

$$r^1 = \frac{6(1000) - (30)(180)}{\sqrt{(6(200) - (30)^2 - (6(5642) - (180)^2}}$$

$$r^1 = \frac{6000) - 5400}{\sqrt{(1200 - (900) - (33852 - 32400)}}$$

$$r^1 = \frac{600}{\sqrt{(300)(1452)}}$$

$$r^1 = \frac{600}{\sqrt{435600}}$$

$$r^1 = \frac{600}{\sqrt{660}} = 0.9091$$

Remarks: They are highly perfect / related.

Illustration: Lasu Campus stores has been selling the believe it or not. Wonders of statistics study guide for 12 Semester and would like to estimates the relationship between sales and number of sections of elementary statistics taught in each Semester. The data below have been collected.

Sales	33	35	24	61	52	45	65	82	29	63	50	79

(units)												
No of sections	3	7	6	6	10	12	12	13	12	13	14	15

a. Obtain the coefficient of correlation
b. Comment on your result?

sales (x)	No. of Section (y)	xy	x^2	y^2
33	3	99	1089	9
38	7	226	1444	49
24	6	144	576	36
61	6	366	3721	36
52	10	520	2704	100
45	12	540	2025	144
65	12	780	4225	144
82	13	1066	6724	169
29	12	348	841	144
63	13	819	3969	169
50	14	700	2500	196
79	15	1185	6241	225

Data

$\sum x = 621$

$\sum y = 123$

$\sum xy = 6833$

$\sum x^2 = 385641$

$\sum y^2 = 15129$

$$r = \frac{\sum xy - (\sum x)(\sum y)}{\sqrt{(\sum x^2 - (\sum x)^2)(\sum y^2 - (\sum y^2)}}$$

$$r = \frac{12\,(6833) - (621)(123)}{\sqrt{(12(385641) - (621)^2 - (12\,(151291) - (123)^2}}$$

$$r = \frac{81996 - 76383}{\sqrt{(432706 - 385641) - (17052 - 15129)}}$$

$$r = \frac{5613}{\sqrt{(47067\,(1923)}}$$

$$r = \frac{5613}{\sqrt{9513.7}} = 0.59$$

Remarks: The sales units and the number of section are particularly correlates or related. The relationship is weak. The unit sales may not necessary be determined or depend on the number of sections.

Illustration: Find the correlation coefficient between the following series. Calculate the correlation of beer consumption as regards the accident in our high ways between 1961 – 1970.

Hence, calculate the dependent variables between bear consumption and road accident.

Year	Road accident	Beer consumption
1961	155	70
1962	150	63
1963	180	72

1964	135	60
1965	156	66
1966	165	70
1967	178	74
1968	160	65
1969	132	62
1970	145	67

Year	Beer consumption (x)	Road accident (y)	xy	x^2	y^2
1961	70	155	10850	4900	24025
1962	63	150	9450	3969	22560
1963	72	180	12960	5084	32400
1964	60	135	5100	3600	18225
1965	66	156	10296	4356	24336
1966	70	165	11760	4900	28224
1967	74	178	13172	5476	31684
1968	65	160	10400	4225	25600
1969	62	132	8184	3844	17424
1970	67	145	9715	4489	21025

Data

V = 10

$\sum x = 669$

$\sum y = 1559$

$\sum xy = 10,4887$

$\sum x^2 = 44943$

$\sum y^2 = 245443$

$$r = \frac{\sum xy - (\sum x)(\sum y)}{\sqrt{(\sum x^2 - (\sum x)^2)(\sum y^2 - (\sum y^2)}}$$

$$r = \frac{10\,(104887) - (660)\,(15590}{\sqrt{(10(44943) - (669)^2 - (10\,(245443) - (1559)^2}}$$

$$r = \frac{1048870 - 1042971}{\sqrt{(449430 - 447651)\,(2454430 - 2436481)}}$$

$$r = \frac{5899}{\sqrt{(779)\,(23949)}}$$

$$r = \frac{5899}{\sqrt{42605271}}$$

$$r = \frac{5899}{6227.27} = 0.9037$$

$= 0.9037$

But $r = 0.8169 = 0.82$

The coefficient determination show the variation in the independent variable (y) as a result of corresponding variation in the explanatory variables (x).

This shows that 90% of beer consumption belong to road accident and of is thus RA = F (BO). The interpretation of coefficient correlation means that 0.82% road accident is brought about 90% of the beer consumption.

4.0 Conclusion

The relationships among business variables can simply be identified using correlation coefficients. Two variables can either be positively or negatively correlated. This correlation can be linear or nonlinear depending on variable characteristics.

5.0 Summary

For a precise quantitative measurement of the degree of correlation between two variables, say X and Y, we use a parameter ⟩ referred to as the correlation coefficient. The sample estimate of this parameter is referred to as r.

6.0 Tutor-Marked Assignment

Determine the correlation between X and Y in the table below.

Period	1	2	3	4	5	6	7	8	9	10
Qty Supply	10	20	50	40	50	60	80	90	90	120
Unit Price (N)	2	4	6	8	10	12	14	16	18	20

7.0 References /Further Reading

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques

UNIT THREE: SPEARMAN'S RANK CORRELATION

RANK CORRELATION OR TIED IN RANK CORRELATION

Contents

1.0 INTRODUCTION

It is found very difficult to quantity a data or set of data that has big values. Rank correlation is used to determine the extent at which the variable are correlated. This idea was employed by Spearman's rank correlation coefficient, which is computed by using this formula.

$$r = \frac{1 - 6\sum d^2}{(x^2 - 1)}$$

2.0 OBJECTIVE

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

- explain the computation of rank correlation coefficients
- apply the concept of correlations in business decisions.

3.0 MAIN CONTENT

Where,

O = number of observation

d = difference between the pairs of rank (x & y) values

r = rank correlation

note: In a cases where there tied or tied in ranking of variables x and y, other representation is applicable.

$$r^1 = \frac{1 - 6 (\sum d^2 + t^3 - t)}{(\pi + 1)(\pi - 1)}$$

Where, t = number of ties in variable x & x respectively, but when coefficient of no determination occur, we equate $1 - r^2$

Illustration: The following data refer to the students scores. The general level of their intelligent in 9 selected courses. Using Spearman's correlated techniques to determine the straight of the relationship between the students cadres and their intelligent.

Sales (units)	y	16	14	15	13	31	16	10	17	20
Intelligent	x	38	41	48	22	64	64	26	53	30

Y	X	rx	ry	d = (rx –	d^2

				ry)	
16	38	6	4.5	1.5	2.25
14	41	5	7	-2	4
15	48	4	6.	-2	4
13	22	9	8	1	1
31	64	1.5	1	0.5	0.25
16	64	1.5	4.5	-3.0	9
10	26	8	9	-1	1
17	53	3	3	0	0
20	30	7	2	5	25

Data

$X = 9$

$\sum F^2 = 46.5$

$r = \frac{1 - 6\sum d^2}{x\ (x^2 - 1)}$

$1 - \frac{6\ (46.5)}{9\ (9^2 - 1)}$

$1 - \frac{6\ (46.5)}{9\ (9^2 - 1)}$

$1 - \frac{279}{720} = 1 - 0.3875 = 0.6125 = 0.61$

Illustrate: A market research asked two (2) smoker to express their difference for 12 difference brands of cigarettes. The reply as shown in the following table.

Brand of cigarette	A	B	C	D	E	F	G	H	I	J	K	L
Smoker z (v)	9	10	4	1	8	11	3	2	5	7	12	6
Smoker W (x)	7	8	3	2	10	12	1	6	5	4	11	9

Requirement: Use Spearman's rank correlation technique to evaluate the straight of relationship between the smokers.

Y	X	rx	ry	d	d^2
9	7	6	4	2	4
10	8	5	3	2	4
4	3	10	9	1	1
1	2	11	12	-1	1
8	10	3	5	-2	4
11	12	1	2	-1	1
3	1	12	10	2	4
2	6	7	11	-4	16
5	5	8	8	0	0
7	4	9	6	3	9
12	11	2	1	1	1
6	9	4	7	-3	9

Data

$\cap = 12$

$\sum d^2 = 54$

$r1 = \underline{1 \cap \sum d^2}$

$n(n^2 - 1)$

$= 1 - \underline{12\ (54)} \qquad 1 - \underline{324}$

$12\ (12^2 - 1) \qquad 1716$

$= 1 - 0.1888$

$r = \quad 0.89$

Illustration: Assuming that 10 men assign to a particular job or task were given two aptitude test. After they have been on the job for some period of time. The production manager was ask to rank the employees from 1st to 10th in regard to their value to the company. You, as the particular manager, should use the Spearman's technique to determine the relationship between the 2 test.

Workers	A	B	C	D	E	F	G	H	I	J
Test 1	96	98	79	78	84	84	76	79	62	44
Test 2	78	72	60	72	64	84	72	56	78	40

Y	X	rx	ry	d	d^2
96	78	2	2.5	-0.15	0.25
98	72	1	4.5	-8.5	12.25
79	60	5.5	7	-2.5	6.25
78	72	7	4.5	2.5	6.25
84	64	3.5	7	-3.5	12.25
84	84	3.5	1	2.5	16.25

76	72	8	4.5	3.5	12.25
79	56	5.5	9	-3.5	12.25
62	78	9	2.5	6.5	42.25
44	40	10	10	0	0

Data

$\cap = 10$

$\sum d^2 = 110.25$

$r1 = \underline{1\ 6\sum d^2}$

$n(n^2 - 1)$

$= 1 - \underline{6 \times 110.25} \qquad 1 - \underline{661.5}$

$10\ (10^2 - 1) \qquad 990$

$= 1 - 0.6682$

$r = \quad 0.3318$

Illustration: The debits in international business transactions (current transfer in million) of United Kingdom from personal sector (x) and central government (y) for the quarters in the period of 1970 to 1972 is given as below:

X	56	57	55	58	51	56	56	58	57	57	57	57
Y	52	40	37	43	57	45	47	51	68	49	43	48

a. Rank in data
b. Compute Spearman's coefficient of rank correlation

X	Y	rx	ry	d	d^2

56	52	9	3	6	36
57	40	5	11	6	36
55	37	11	12	1	1
58	43	1.5	9.5	8	64
51	57	1.2	2	10	100
56	45	9	8	1	1
56	47	9	7	2	4
58	51	1.5	4	2.5	6.25
57	68	5	1	4	16
57	49	5	5	0	0
57	43	5	9.5	4.5	20.25
57	48	5	6	1	1

Data

$\cap = 12$

$\sum d^2 = 285.5$

$r1 = \underline{1\ 6\sum d^2}$

$n(n^2 - 1)$

$= 1 - \underline{6 \times 285.5} \qquad 1 - \underline{1713}$

$12\ (12^2 - 1) \qquad 1716$

$= 1 - 0.9983$

$r = 0.0001$

Comment: The value of r1 shows that x and y are not correlated i.e. they are not in agreement

3.1 ANALYSIS OF RANKED DATA

Spearman's coefficient of correlation assumes the data to be at least interval scale. Chalse – Spearman, a British statistician, introduced a measure of correlation for ordinal- level data known as Spearman's rank-order correlation coefficient (i.e. A measure of relationship between two sets of ranked data). Its designated as r_s, may range between -1.0 and +1.0 inclusive with -1.0 and +1.0 representing perfect rank correlation. Zero indicates no rank correlation.

The general representation can be given as

$$r_s = 1 - \frac{6\ (\sum d2)}{n\ (n^2 - 1)}$$

where,

n = number of paired observations

d = difference between the ranks for each pair.

NB: for large-sample where n is 10 or more, the student's "t" distribution can be used as the test of statistic. And the degree of freedom is given as (n -2)

The general computed formular is given as

$$t = r_s \quad \frac{\pi - 2}{1 - r_s^2}$$

Example: A sample of 12 auto mechanics was ranked by the supervisor regarding their mechanical ability and their social compatibility. The results are as follows:

Worker	Mechanical Ability	Social compatibility
1	1	4

2	2	3
3	3	2
4	4	6
5	5	1
6	6	5
7	7	8
8	8	12
9	9	11
10	10	9
11	11	7
12	12	10

Compute the coefficient of rank correlation can we conclude that there is a positive association in the population between the ranks of mechanical ability and social compatibility?

Use the 0.05 significance level.

Worker	Mechanical Ability	Social compatibility	d	d^2
1	1	4	-3	9
2	2	3	-1	1
3	3	2	1	1
4	4	6	-2	4
5	5	1	4	16
6	6	5	1	1

7	7	8	-1	1
8	8	12	-4	16
9	9	11	-2	4
10	10	9	1	1
11	11	7	4	16
12	12	10	2	4

$\sum d^2 = d\ 4$

$r_s = \frac{6\sum d^2}{(n^2 - 1)}$

$1 - \frac{6(74)}{12(12-1)} = 1 = 0.741$

Decision: The value 0.741 indicate fairly strong positive association between the ranks of mechanical ability and social compatibility.

$\alpha = 0.05$

H_0: The rank correlation in the population is zero

H_1: the rank correlation in the population is greater than zero.

Using one tailed test.

d.f = n -2 = 12 – 2 = 10

To obtain “t” test

$t = rs\sqrt{\frac{n-2}{1-r^2_s}}$

$= 0.741\ \sqrt{\frac{12-2}{1-(0.74)^2}}$ $0.741\ \sqrt{\frac{10}{1-(0.74)^2}}$

0.741 (22.177) = 0.0741 $\sqrt{(4.7092)}$

= 3.4895 = 3.49

Decision: Since computer value exceed critical value of 1.812, then H_0 is rejected, and H_1 is accepted. It is concluded that there is a positive association between the ranks of social compatibility and mechanical ability among auto mechanics.

4.0 Summary and Conclusion

5.0 Tutor-Marked Assignment

1. Twelve persons whose IQs were measured in cottage between 1960 and 1965 were located recently and retested with an equivalent IQ test. The information is given below.

Student	Recent score	Original score
John Barr	119	112
Bill Sedwick	103	108
Morica Elephant	115	115
Ginge Tale	109	100
Larry Clark	131	120
Jim Redding	110	108
Carol Papalia	109	113
Victor Soppa	113	126
Dallae Paul	94	95
Carol Kozoloski	119	110
Jok Sass	118	117

P.S Sundar	112	102

At the 0.05 significance level can we conclude that the IQ scores have increased in over 20 years. Compute the coefficient of rank correlation.

6.0 Further Reading

NOUN TEXT BOOK, ENT 321: Quantitative Methods for Business Decisions

7.0 References/ Further Reading

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques

UNIT 4: LEAST SQUARE REGRESSION ANALYSIS

Contents

1.0 INTRODUCTION

Regression analysis can be defined as the relationship between two or more variables. This relationship has to do with the changes that result from a change in one of the related variables.

2.0 OBJECTIVE

The main objective of this unit is to enable students understand the theory behind and the application of regression analysis in statistics. Students are expected at the end of this unit to be able to apply regression analysis to solving day-to-day business and economic problems.

3.0 MAIN CONTENT

Uses and the Types of Regression

i. It is used for prediction
ii. It is used for description of relationship
iii. To improve on knowledge of variable of interest

Basically, there are two types:

i. Simple (linear) regression
ii. Multiple (non linear) regression

Simple (linear) regression

This involve only two variables and the relationship between them tends towards a fixed direction.

Multiple (non linear) regression

This also involved more than two variables in the regression model or equation.

Mathematically, let us assume that x_1 and x_2 as independent variable (factor) and y as dependent variable. The independent variable may be more than two, i.e. rt can be obtainable as x_1, x_2, x_3….x

Recall, y = a + bx = simple regression

Similarly, we can have $y = a + b_1x_1 + b_2x_2$ (for multiple regression)

Method of Calculating Regression Line

Regression line of any form can be titled to a bivariate data by any of the following methods.

1. Freehand method

In this method, regression line is fitted into the scatter diagram

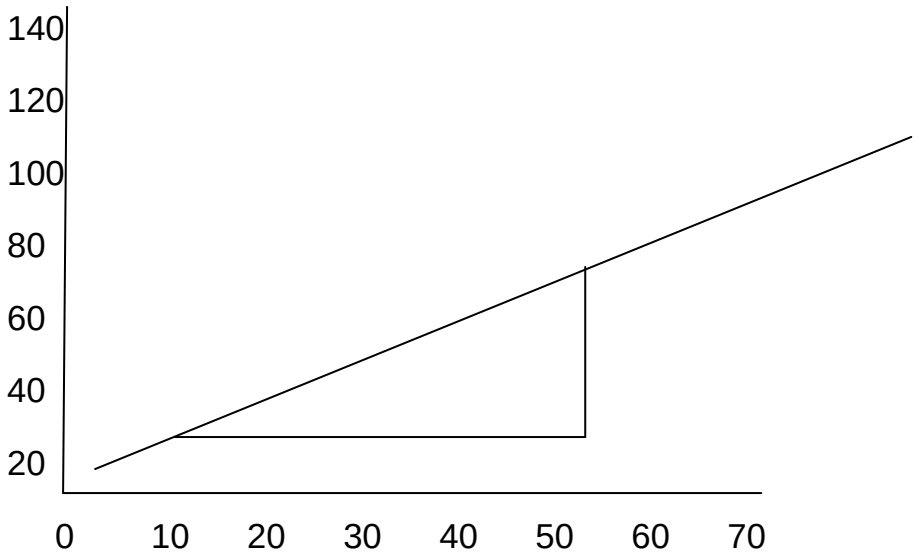
This scatter diagram is the graphically representation of relationship which exists between two variables by drawing a lines of best fit through the various points which are estimate from the relationship x and y.

Illustration: Given/estimate the regression equation by using the scatter diagram from the data below. The marks scored by a group of philosophy students and mathematics students are as follows.

Philosophy marks	38	51	19	53	39	38	66

Mathematics marks	50	32	36	54	52	56	80

By scatter diagram:



Limitation:

i. It does not give a unique regression line
ii. It also does not give unique regression coefficient

2. Least Square Method
   This is the mathematical method of determined the points estimate of ‘a’ and ‘b’ from the available sample points. This method is the most reliable of all the methods. A gives a unique regression line and a unique regression coefficient. The method of least square provides two

set of equation called (Normal Equations) which can solved simultaneously for two unknown.

By representation,
Y = a+bx; b = the coefficient of x and x = independent variable.

From the line of fit from y = a + bx
$\sum y = an + b\sum x$----------i
$\sum xy = a\sum x + b\sum x^2$ -----ii

Both equation are regretted a normal equation.
$ax + b\sum x$
$a\sum x + b\sum x^2$

From the equation above,
$ax + b\sum x - \sum y$
$a\sum x + b\sum x^2 = \sum xy$
by determinant method

$$\begin{Bmatrix} x & \sum x \\ \sum x & \sum x^2 \end{Bmatrix} \quad \begin{Bmatrix} a \\ b \end{Bmatrix} = \begin{Bmatrix} \sum y \\ \sum xy \end{Bmatrix}$$

$$\begin{Bmatrix} x & \sum x \\ \sum x & \sum x^2 \end{Bmatrix}$$

To obtain

$$\begin{Bmatrix} x & \sum x \\ \sum x & \sum x^2 \end{Bmatrix} = N \sum x^2 - \sum x \cdot \sum x$$
$$= \sum x^2 - \sum x^2$$

$= \sum x^2 - (\sum x)^2$

$(A) = \sum x^2 - (\sum x)^2$ -------3

From (A_1) $\begin{Bmatrix} x & \sum x \\ \sum xy & \sum x^2 \end{Bmatrix}$ $= (A1) = (\sum x^2 \cdot \sum y - \sum x \cdot \sum xy)$

$(A^1) = \sum x^2 . \sum y - \sum x . \sum xy$ -------4

From other Equation

(A_2) $\begin{Bmatrix} x & \sum x \\ \sum x & \sum xy \end{Bmatrix}$ $= \sum xy - \sum x . \sum y$ ----------5

Mathematically, (By determinant)

$(A1) = \underline{\Delta A^1} . \underline{\Delta A^2}$

$\Delta 1 \quad \Delta$

$= \underline{\sum x^2 . \sum y - \sum x . \sum xy}$……a

$\sum x^2 – (\sum x)^2$

$(A2) = \underline{\sum x^2 . \sum y - \sum x . \sum xy}$ ----------b

$\sum x^2 – (\sum x)^2$

Both equation should be memorized.

Non-Linear Model

On most occasion, the simple linear model and in particular the multiple linear model will not be satisfactory. A plot or scatter diagram on the

dominant variable may suggest that the relationship is not linear. We consider non-linear model, which involves:

i. Different type of curve
ii. Linearization

Types of Curve

There are 3 main types of curve

1. Exponential (Growth) curve: This is a situation whereby when a data is expected to grow by some proportion or percentage in each period.
   An exponential curve have:
   $Y = ab^{ru}$ and in particular or
   $Y = ac^{ru}$ or ab^{u}
   When abr are constant
   Where: y = variable to be predicted
   Axb = constant
   X = number of period
   Now, to linearise the above equation
   $Y = ab^{u}$
   Obtain log of both sides
   $\text{Log } y = \log A + \log b^{x}$
   $\text{Log } y = \log A + x\log B$
   Equate $\log_u$ to both sizes
   Log y = A + Bx ----------x

2. Hyperbolic Model (curve): This has a formular
   Y = a + b/x or y = 1/a+bx
   To linearise y, take x = 1/x, then we have
   Y = a +bx
   i/y = a + bx
   since y = 1/y
   therefore, y = a +bx
3. Power curve mor model: This power model have the form of y = axb. Otherwise known as logarithms functions. The general representation can be given as:
   $y = ax^{b}$

to linearise: obtain $\log_{10}$ to both sides
$\log_{10} y = \log_{10} a + b\log_{10} x$
since a and b are constant.
$\text{Log}_{10} y = y^1$ and $\log_{10} a = A$
Therefore, $y^1 = A + bx$

Illustration: Draw the scatter diagram and fit an exponential curbe in the following data

Years	1983	1984	1985	1986	1987
Sales	100	150	225	337.5	506.25

X years	Y sales
O	100
1	150
2	225
3	337.5
	506.25

X	Y	Log y	Xlogy	X^2
0	100	2.000	0	0
1	150	2.1761	2.1761	1
2	225	2.3522	4.7044	4
3	337.5	2.5282	7.5846	9
4	506.25	2.7045	10.8180	16

Data

$\sum x = 10$

$\sum y = 11.7610$

$\sum xy = 25.2831$

$\sum x^2 = 30$

From general representation

$y = abx$

$\log y = \log a + x \log b$

to find a and b

firstly, to find b = ?

$$b = \frac{n\sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$b = \frac{5(25.283) - (10)(11.7610)}{5(30) - 100}$$

$$b = \frac{8.8056}{50} = 0.17611$$

Then, $\log b = 0.17611$

$b^{-1} (0.17611) = 1.5$

to obtain a = ?

$A = y - bx$

$A = y - bx$

$$A = \frac{\sum y}{n} - b\frac{\sum x}{n}$$

$$a = \frac{12.6532}{5} - \frac{(2)(2.0792)}{5}$$

A = 1.6989

A^{-1} (1.6989) = 50

Therefore, $y = ax^b$

$Y = 50(x^2)$

But when x = 1, 2, 3, 4, -----

Therefore, $y = 50 (1^2) = 5$

$Y = 50 (2^2) = 200$

$Y = 50 (3^2) = 450$

4.0 Conclusion

The relationships among business variables can simply be identified using correlation coefficients. Two variables can either be positively or negatively correlated. This correlation can be linear or nonlinear depending on variable characteristics.

5.0 Summary

For a precise quantitative measurement of the degree of correlation between two variables, say X and Y, we use a parameter ⟩ referred to as the correlation coefficient. The sample estimate of this parameter is referred to as r.

6.0 Tutor-Marked Assignment

1. Illustration: Given/estimate the regression equation by using the scatter diagram from the data below. The marks scored by a group of philosophy students and mathematics students are as follows.

Philosophy marks	3	5	9	5	9	8	6
Mathematics marks	5	2	3	4	5	6	8

2. Illustration: Draw the scatter diagram and fit an exponential curbe in the following data

Years	1983	1984	1985	1986	1987

Sales	10	15	25	33.75	50.625

7.0 References/Further Reading

NOUN TEXT BOOK, ENT 321: Quantitative Methods for Business Decisions

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques

UNIT 5: MULTIPLE REGRESSION ANALYSIS

Content

1.0 INTRODUCTION

Recall, the degree of relationship that connect three or more variables together are called multiple correlation regression.

e.g. $y = a+b_1x_1+b_2x_2+b_3x_3\text{------}b_x\ x_n$

2.0 OBJECTIVE

The objective of this unit is to introduce students to multiple regression analysis and emphasize its applications in statistics.

3.0 MAIN CONTENT

The above expression can be solved by the normal equation of the three variables.

$y = a+b_1x_1+b_2x_2+b_3x_3\text{------}b_x\ x_n$

$\sum y = ax + b_1\sum x_1+b_2\sum x_2\text{----------(i)}$

$\sum x_1y = a\sum x_1+b_1\sum x_1^2+b_2\sum x_1x_2\text{----------(ii)}$

$\sum x_2y = a\sum x_2+b_1\sum x_1^2,x_2+b_2\sum x^2{}_2\text{--------(iii)}$

But the coefficient of multiple determination r^2 can be expressed as:

$$r^2 = \frac{a\sum y+b_1\sum x_1y+b_2\sum x_2y - (\sum y)^2}{\sum y^2 - (\sum y/x)^2}$$

Illustration: The Faculty of Management Science (HMS) has investigating the relationship between some students performance in their various courses and lecture received per each semester and also the quality of some lecturers. The faculty has a data of ten candidates which are:

Student	1	2	3	4	5	6	7	8	9	10
No of lecturer	9	6	12	14	11	6	19	16	3	9
Quality of lecturers	99	100	119	95	110	117	98	101	100	115
Exams scores	56	45	80	73	71	55	95	86	34	66

From general representation

$y = a+b_1x_1+b_2x_2+b_3x_3$------$b_x\ x_n$

Students	y	y^2	x_1	x_1^2	x_2	x^2_2	x_1y	x_2y	x_1x_2
1	56	3136	9	81	99	9801	5044	5544	891
2	45	2025	6	36	100	10,000	270	4500	600
3	50	6400	12	144	119	14161	960	9520	1428
4	73	5329	14	196	95	9025	1022	6935	1330
5	71	5041	11	121	110	12100	781	7810	1210
6	55	3025	6	36	117	13689	330	6435	702
7	95	9025	19	361	98	9604	1805	9310	1862
8	86	7396	16	256	101	10201	1376	8656	1616

9	34	1156	3	9	100	10000	102	3400	300
10	66	4356	9	81	115	13225	594	7590	1035

Data:

$\sum y = 661$

$\sum y^2 = 46889$

$\sum x_1 = 105$

$\sum x_1^2 = 1321$

$\sum x2 = 1054$

$\sum x_2^2 = 111{,}806$

$\sum x_1 y = 7744$

$\sum x_2 y = 69730$

$\sum x_1 x_2 = 10{,}974$

To find b = p

$$bx_1 = \frac{n \sum x_1 y - \sum x_1 y}{n \sum x_1^2 - (\sum x_1)^2}$$

$$= \frac{10\,(7744) - (105)\,(661)}{10\,(1321) - (105)^2}$$

$$= \frac{77440 - 69405}{13210 - 11025} = \frac{8035}{2185} = 3.6773 = 3.68$$

Therefore: bx = 3.68

To find a = ?

$$ax_1 = y - \frac{bx1 \sum x1}{n}$$

$$ax1 = \frac{\sum y}{n} - \frac{bx_1 \sum x_1}{n}$$

$$= \frac{661}{10} - \frac{(3.68)(105)}{10}$$

$$= 66.1 - 38.64 = 27.46$$

Therefore, $ax_1 = 27.64$

But, $y = ax_1 + bx_1x_1$

$y = 27.64 + 3.68$

$y = 27.64 + 3.68x_1$

4.0 Conclusion

The relationships among business variables can simply be identified using correlation coefficients. Two variables can either be positively or negatively correlated. This correlation can be linear or nonlinear depending on variable characteristics.

5.0 Summary

For a precise quantitative measurement of the degree of correlation between two variables, say X and Y, we use a parameter referred to as the correlation coefficient. The sample estimate of this parameter is referred to as r.
A **partial correlation coefficient** measures the relationship between any two variables, keeping other variables constant.
The limitations of linear correlations as a technique for the study of economic relations are as follows

The formula for correlation coefficient applies only to linear relationships between variables. That correlation coefficient as a measure of co-variability of variables does not imply any functional relationship between the variables concerned.

6.0 Tutor Mark Assignment

Illustration: calculate the coefficient of linear multiple regression of the data below: the association of accountants is investigating the relationship between performance in Quantitative methods and how studied per week

and the general level of intelligence of candidates. The Association has data on ten students which are:

Students	1	2	3	4	5	6	7	8	9	10
Hours studied x_1	9	6	12	14	11	6	19	16	3	9
T.Q (x_2)	99	100	119	95	110	117	98	101	100	115

Hence, predict the expected score of a candidate.

7.0 References/Further Reading

NOUN TEXT BOOK, ENT 321: Quantitative Methods for Business Decisions

OTOKOTI O.S. Contemporary Statistics

JIDE JONGBO Fundamental Statistics for Business

KEHINDE J.S. Statistics Method & Quantitative Techniques.

MODULE FIVE: STATISTICAL TEST

UNIT 1:

T–TEST

CONTENTS

1.0 INTRODUCTION

The Student's t-test

For large sample test for mean $Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1), asymptotically$

If the population variance is unknown then for the large samples, its estimates provided by sample variance S^2 is used and normal test is applied. For small samples an unbiased estimate of population variance σ^2 is given by:

$$S^2 = \frac{1}{n-1}\sum(\bar{x} - x)^2 \rightarrow ns^2 = (n-1)S^2$$

It is quite conventional to replace σ^2 byS^2 (for small samples) and then apply the normal test even for small samples. W.S Goset, who wrote under the pen name of Student, obtained the sampling distribution of the statistic $\frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}}$for small samples and showed that it is far from normality. This discovery started a new field, viz '*Exact Sample Test*' in the history of statistical inference.

Note: If $x_1, x_2 \ldots\ldots\ldots\ldots x_n$ is a random sample of size n from a normal population with mean μ and variance σ^2 then the Student's t statistic is defined as:

$$t = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} = \frac{\bar{x}-\mu}{\sqrt{\frac{S^2}{n}}}$$

Where $\bar{x} = \frac{\Sigma x}{n}$ is the sample mean and $S^2 = \frac{1}{n-1}\sum(\bar{x}-x)^2$ is an unbiased estimate of the population variance σ^2

2.0 OBJECTIVES

The objective of this unit is to introduce students to *t*-distribution and emphasize its application in statistics.

3.0 MAIN CONTENT

3.1 Applications of *t*-distribution

(i) *t*-test for the significance of single mean, population variance being unknown

(ii) *t*-test for the significance of the difference between two sample means, the population variances being equal but unknown

(iii)*t*-test for the significance of an observed sample correlation coefficient

3.2 Test for Single Mean

Sometimes, we may be interested in testing if:

(i) The given normal population has a specified value of the population mean, say μ_o.

(ii) Thesample mean$\bar{x}$ differ significantly from specified value of population mean.

(iii)A given random sample $x_1, x_2 \ldots\ldots\ldots\ldots x_n$of size n has been drawn from a normal population with specified meanμ_o.

Basically, all the three problems are the same. We set up the corresponding null hypothesis thus:

(a) H_o: $\mu = \mu_o$i.e the population mean is μ_o

(b) H_o: There is no significant difference between the sample mean and the population mean. In order words, the difference between $\bar{x}$ and μ is due to fluctuations of sampling.

(c) H_o: The given random sample has been drawn from the normal population with mean μ_o. Under H_o the test-statistic is:

$$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{\bar{x} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$$

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

And it follows Student's t-distribution with (n-1) degrees of freedom.

We compute the test-statistic using the formula above under H_o and compare it with the tabulated value of *t* for (n-1) *d.f.*at the given level of significance. If the absolute value of the calculated *t* is greater than tabulated *t*, we say it is significant and the null hypothesis is rejected. But if the calculated *t* is less than tabulated *t*, H_o may be accepted at the level of significance adopted.

3.3 Assumptions for Student's test

(i) The parent population from which the sample is drawn is normal

(ii) The sample observations are independent i.e. the given sample is random.

(iii)The population standard deviation σ is unknown

Example: Ten cartons are taken at random from an automatic filling machine. The mean net weight of the 10 cartons is 11.8kg and standard deviation is 0.15kg. Does the sample mean differ significantly from the intended weight of 12kg, α=0.05

Hint: You are given that for *d.f.* =9, $t_{0.05}$ = 2.26

Solution: n= 10,$\bar{x}$= 11.8kg, s = 0.15kg

Null hypothesis, H_o: μ = 12 kg (i.e. the sample mean of $\bar{x}$ =11.8 kg does not differ significantly from the population mean μ = 12 kg

*Alternative Hypothesis.*Ho: ≠ 12kg (Two tailed)

$$t = \frac{\bar{x}-\mu}{\sqrt{\frac{S^2}{n}}} = \frac{\bar{x}-\mu}{\sqrt{\frac{S^2}{n-1}}} \sim t_{n-1} = t_9$$

$$t = \frac{11.8-12}{0.15\sqrt{9}} = \frac{-0.2\,X3}{0.15} - 4.0$$

The tabulated value of t for 9 d.f. at 5% level of significance is 2.26. Since the calculated t is much greater than the tabulated t, it is highly significant. Hence, null hypothesis is rejected at 5% level of significance and we conclude that the sample mean differ significantly.

3.4 t-Test for difference of means

Assume we are interested in testing if two independent samples have been drawn from two normal populations having the same means, the population variances being equal.

Let $x_1, x_2 \ldots\ldots\ldots\ldots x_n$ and $y_1, y_2 \ldots\ldots\ldots\ldots y_n$ be two independent random samples from the given normal populations.

H_o: $\mu_x = \mu_y$i.e the two samples have been drawn from the normal populations with the same means. Under the hypothesis that the $\sigma_1^2 = \sigma_2^2 = \sigma^2$i.e population variances are equal but unknown, the test statistic under Ho is:

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Where $\bar{x} = \frac{1}{n_1}\sum x, \quad \bar{y} = \frac{1}{n_2}\sum y$

And $S^2 = \frac{1}{n_1+n_2-2}[\sum(\bar{x} - x)^2 + \sum(\bar{y} - y)^2]$

This is an unbiased estimate of the common population variance σ^2 based on both the samples. By comparing the computed value of t with the tabulated value of t for $n_1 + n_2 - 2$ *d.f.* and at desired level of significance, usually 5% or 1%, we reject the null hypothesis.

Example: The nicotine content in milligram of two samples of tobacco were found to be as follows:

Sample A:	24	27	26	21	25	
Sample B:	27	30	28	31	22	36

Can it be said that the two samples come from the same normal population having the same mean?

Solution Hints: Applying the above formula and calculating the variance as appropriate, the calculated t-value is -1.92. the tabulated value for 9 d.f. at 5% level of significance for two-tailed test is 2.262. Since calculated t is less than the tabulated t, it is not significant and the null hypothesis is accepted.

4.0 CONCLUSION

T-testhas very wide applications. It can be applied in the tests of single mean, in the comparison of two different means and in the test of significance of other parameter estimates.

5.0 SUMMARY

Here, you would have learnt how to apply t-test in solving statistical problems such as test to confirm if mean is a certain value, to test significance of the difference between two mean among others.

6.0 TUTOR-MARKED ASSIGNMENT

1. The mean weekly sale of the chocolate bar in candy stores was 146.3 bars per store. After advertising campaign the mean weekly sales in 22 stores for typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful?
2. Prices of shares of a company on the different days in a month were found to be: 66, 65, 69, 70, 69, 71, 70, 63, 64 and 68. Discuss whether the mean price of the price of the shares in the month is 65.

3.Two salesmen A and B are working in certain district. From a Sample Survey Conducted by the Head Office the following results were obtained. State whether there is any significant difference in the average sales between the two salesmen.

	A	B
No. of sales	20	18
Average sales (in '000₦)	170	205
Average sales (in '000₦)	20	25

7.0 REFERENCES / FURTHER READING

Spiegel, M. R., Stephens L.J., (2008).*Statistics*.(4th ed.). New York, McGraw Hill press.

Gupta S.C. (2011). *Fundamentals of Statistics*.(6th Rev.& Enlarged ed.).Mumbai India, Himalayan Publishing House.

Swift L., (1997).*Mathematics and Statistics for Business, Management and Finance.*London UK, Macmillan.

UNIT 2:

F-TEST

CONTENTS

1.0 INTRODUCTION

In F-TEST, If X is a χ^2-variate with n_1 degree of freedom and Y is an independent χ^2-variate with n_2 degree of freedom, then *F*-statistic is defined as:

$$F = \frac{X/n_1}{Y/n_2}$$

i.e. F-statistic is the ratio of two independent chi-square variates divided by their respective degrees of freedom. The statistic follows G.W Snedecor's*F*-distribution with (n_1, n_2)degree of freedom with probability density function given by:

$$(F) = y_o \cdot \frac{F^{\frac{n_1}{2}-1}}{(1+\frac{n_1}{n_2}F)^{\frac{n_1+n_2}{2}}} \; ; 0 \leq F < \infty$$

Where y_o is a constant which is so determined that total area under the probability curves is 1 i.e $\int_0^\infty p(F)\,dF = 1.$ *This gives :* $y_o = \frac{(\frac{n_1}{n_2})^{n\ 1/2}}{\beta(\frac{n\ 1}{2}, \frac{n\ 2}{2})}$

Note: The sampling distribution of F-statistics does not involve any population parameters and depends only on the degrees of freedom n_1 and n_2. The graph of the function $p(F)$ varies with the degree of freedom n_1 and n_2.

Critical values of F-distribution: The available *F*-tables in most standard statistical table give the critical values of *F* for the right-tailed test, i.e. the critical region is determined by the right tail areas. Thus, the significant value $F\alpha\ (v_1,v_2)$ at level of significance α and (v_1,v_2) d.f. is determined by the equation:

$$P\ [F > F\alpha\ (v_1, v_2)] = \alpha$$

Significant values of the variance-Ratio $F = \frac{S_1^2}{S_2^2}\ ;\ S_1^2 > S_2^2$

2.0 OBJECTIVE

The main objective of this section is to introduce student to the world of F-distribution and learn its theories and application to day-to-day business and economic problems.

3.0 MAIN CONTENT

3.1 Applications of the F-distribution

F-distribution has a number of applications in the field of statistics. This includes but not limited to the following:

(1) To test for equality of population variances

(2) To the equality of several population means i.e. for testing Ho: $\mu_1 = \mu_2 = ... = \mu_k$. This is by far the most important application of F-statistic and is done through the technique of Analysis of Variance (ANOVA). This shall be treated as a separate unit later.

(3) For testing the significance of an observed sample multiple correlation

(4) For testing the significance of an observed sample correlation ratio

3.2 For testing equality of population variances: Here, we set up the Null hypothesis Ho: $\sigma_1 = \sigma_2 = \sigma$,i.e. population variances are the same. In other words, H_o is that the two independent estimates of the common population variance do not differ significantly.

Under Ho, the test statistic is

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1),$$

Where, S_1^2 and S_2^2 are unbiased estimates of the common population variance σ^2 and are given by:

$$S_1^2 = \frac{1}{n_1 - 1}\sum(x - \bar{x})^2 \quad and \quad S_2^2 = \frac{1}{n_2 - 1}\sum(y - \bar{y})^2$$

and it follows Snedecor's F-distribution with $v_1 = n_1 - 1$, $v_2 = n_2 - 1$ *d.f.; i.e.* $F \sim F(v_1, v_2)$

Since F-test is based on the ratio of two variances, it is also known as variance ratio test.

Assumption for F-test for equality of variances

1. The samples are simple random samples
2. The samples are independent of each other
3. The parent populations from which the samples are drawn are normal

N.B (1) since, the most available tables of the significant values of F are for the right-tail test, i.e. against the alternative H_o: $\sigma_1^2 > \sigma_2^2$, in numerical problems we will tale greater of the variances S_1^2 or S_2^2 as the numerator and adjust for the degree of freedom accordingly. Thus, in $F \sim (v_1, v_2)$, v_1 refers to the degree of freedom of the larger variance, which must be taken as the numerator while computing *F*.

If Ho is true i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ the value of F should be around 1, otherwise, it should be greater than 1. If the value of *F* is far greater than 1 the Ho should be rejected. Finally, if we take larger of S_1^2 or S_2^2 as the numerator, all the tests based on the F-statistic become right tailed tests.

- All one tailed tests for Ho at level of significance "α" will be right tailed tests only with area "α" in the right.
- For two-tailed tests, the critical valuesare located in the right tail of F-distribution with area ($\alpha/_2$) in the right tail.

Example 1: The time taken (in minutes) by drivers to drive from Town A to Town B driving two different types of cars X and Y is given below

Car Type X:	20	16	26	27	23	22	
Car Type Y:	27	33	42	35	32	34	38

Do the data show that the variances of time distribution from population from which the samples are drawn do not differ significantly?

Solution:

X	***d = x – 22***	d^2	***Y***	***d = y -35***	D^2
20	-2	4	27	-8	64
16	-6	36	33	-2	4
26	4	16	42	7	49
25	5	9	35	0	0
23	1	1	32	-3	9
22	0	0	34	-1	1
			38	3	9
Total	**2**	d^2 **= 82**		**-4**	ΣD^2 **=136**

$$S_1^2 = \frac{1}{n_1-1}\sum(x - \bar{x})^2 = \frac{1}{n_1-1}\left[\sum d^2 - \frac{(\sum d)^2}{n_1}\right]$$

$$= \frac{1}{5}\left[82 - \frac{4}{6}\right] = \frac{1}{5}[82 - 0.67] = 16.266$$

$$S_2^2 = \frac{1}{n_2-1}\sum(y - \bar{y})^2 = \frac{1}{n_2-1}\left[\sum D^2 - \frac{(\sum d)^2}{n_1}\right]$$

$$= \frac{1}{6}\left[136 - \frac{16}{7}\right] = \frac{1}{6}[136 - 2.286] = 22.286$$

Since, $S_2^2 > S_1^2$, under Ho, the test statistic is

$$F = \frac{S_2^2}{S_1^2} \sim F(n_1 - 1, n_2 - 1) = F(6,5)$$

$$F = \frac{22.286}{16.266} = 1.37$$

Tabulated $F_{0.05(6,5)} = 4.95$

Since the calculated F is less than tabulated F, it is not significant. Hence Ho may be accepted at 5% level of significance or risk level. We may therefore conclude that variability of the time distribution in the two populations is same.

4.0 CONCLUSION

In conclusion, F-test can be used to test the equality of several population variances, several population means, and overall significance of a regression model.

5.0 SUMMARY

Students have learnt the theories and application of the F-test

6.0 TUTOR-MARKED ASSIGNMENT

Can the following two samples be regarded as coming from the same normal population?

Sample	Size	Sample Mean	Sum of squares of deviation from the mean
1	10	12	120
2	12	15	314

7.0 REFERENCE/FURTHER READING

Spiegel, M. R., Stephens L.J., (2008).*Statistics*.(4th ed.). New York, McGraw Hill press.

Swift L., (1997).*Mathematics and Statistics for Business, Management and Finance.*London UK, Macmillan.

UNIT 3:

CHI-SQUARE TEST

CONTENTS

1.0 INTRODCUTION

The square of a standard normal variable is called a Chi-square variate with 1 degree of freedom, abbreviated as *d.f*. Thus if x is a random variable following normal distribution with mean μ and standard deviation σ, then $(X- \mu)/ \sigma$ is a standard normal variate.

Therefore, $Z = (\frac{x-\mu}{\sigma})^2$ is a chi-square (abbreviated by the letter χ^2 of the Greek alphabet) variate with 1 *d.f.*

If X_1, X_2, X_3, ...X_v are v independent random variables following normal distribution with means μ_1, μ_2, μ_3,..... μ_v, and standard deviations σ_1, σ_2, σ_3,..... σ_v respectively then the variate

$$\chi^2 = (\frac{x_1-\mu_1}{\sigma 1})^2 + (\frac{x_2-\mu_2}{\sigma 2})^2 + \cdots \ldots \ldots (\frac{x_v-\mu_v}{\sigma v})^2$$

$$= \sum_{i=1}^{v} (\frac{x_1-\mu_1}{\sigma 1})^2$$

which is the sum of the squares of v independent standard normal variates, follow Chi-square distribution with v *d.f.*

7.1 OBJECTIVE

The main objective of this unit is to enable students understand the theory behind and the application of chi-square statistics. Students are expected at the end of this unit to be able to apply chi-square analysis to solving day-to-day business and economic problems.

3.0 MAIN CONTENTS

3.1 Applications of the χ^2-Distribution

Chi-square distribution has a number of applications, some of which are enumerated below:

(i) Chi-square test of goodness of fit.

(ii) χ^2-test for independence of attributes

(iii)To test if the population has a specified value of variance σ^2.

(iv)To test the equality of several population proportions

Observed and Theoretical Frequencies

Suppose that in a particular sample a set of possible events E_1, E_2, E_3,.....E_k are observed to occur with frequencies O_1, O_2, O_3, ...O_k, called observed frequencies, and that according to probability rules they are expected to occur with frequencies e_1, e_2, e_3,.....e_k, called expected or theoretical frequencies. Often we wish to know whether the observed frequencies differ significantly from expected frequencies.

Definition of χ^2

A measure of discrepancy existing between the observed and expected frequencies is supplied by the statistics χ^2 given by

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \cdots \cdots \cdots \cdots \cdots \cdots \frac{(o_k - e_k)^2}{e_k}$$

3.2 Chi-Square test of goodness of fit

The chi-square test can be used to determine how well theoretical distributions 9such as the normal and binomial distributions) fit empirical distributions (i.e. those obtained from

sample data). Suppose we are given a set of observed frequencies obtained under some experiment and we want to test if the experimental results support a particular hypothesis or theory. Karl Pearson in 1900, developed a test for testing the significance of the discrepancy between experimental values and the theoretical values obtained under some theory or hypothesis. This test is known as χ^2-test of goodness of fit and is used to test if the deviation between observation (experiment) and theory may be attributed to chance (fluctuations of sampling) or if it is really due to the inadequacy of the theory to fit the observed data.

Under the null hypothesis that there is no significant difference between the observed (experimental and the theoretical or hypothetical values i.e. there is good compatibility between theory and experiment.

Karl Pearson proved that the statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots \ldots\ldots\ldots\ldots\ldots \frac{(O_n - E_n)^2}{E_n}$$

Follows χ^2-distribution with $v = n-1$, *d.f.* where $O_1, O_2, \ldots\ldots\ldots\ldots O_n$ are the observed frequencies and $E_1, E_2, \ldots\ldots\ldots\ldots E_n$ are the corresponding expected or theoretical frequencies obtained under some theory or hypothesis.

3.3 Steps for computing χ^2 and drawing conclusions

(i) Compute the expected frequencies $E_1, E_2, \ldots\ldots\ldots\ldots E_n$ corresponding to the observed frequencies $O_1, O_2, \ldots\ldots\ldots\ldots O_n$ under some theory or hypothesis

(ii) Compute the deviations $(O-E)$ for each frequency and then square them to obtain $(O-E)^2$.

(iii) Divide the square of the deviations $(O-E)^2$ by the corresponding expected frequency to obtain $(O-E)^2/E$.

(iv) Add values obtained in step (iii) to compute $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

(v) Under the null hypothesis that the theory fits the data well, the statistic follows χ^2-distribution with v = n-1 *d.f.*

(vi) Look for the tabulated (critical) values of χ^2 for (n-1) d.f. at certain level of significance, usually 5% or 1%, from any Chi-square distribution table.

If calculated value of χ^2 obtained in step (iv) is less than the corresponding tabulated value obtained in step (vi), then it is said to be non-significant at the required level of significance. This implies that the discrepancy between observed values (experiment) and the expected values (theory) may be attributed to chance, i.e. fluctuations of sampling. In other words, data do not provide us any evidence against the null hypothesis [given in step (v)] which may, therefore, be accepted at the required level of significance and we may conclude that there is good correspondence (fit) between theory and experiment.

(vii) On the other hand, if calculated value of χ^2 is greater than the tabulated value, it is said to be significant. In other words, discrepancy between observed and expected frequencies cannot be attributed to chance and we reject the null hypothesis. Thus, we conclude that the experiment does not support the theory.

Example 1: A pair of dice is rolled 500 times with the sums in the table below

Sum (x)	**Observed Frequency**
2	15
3	35
4	49
5	58
6	65
7	76
8	72
9	60
10	35
11	29
12	6

Take $\alpha = 5\%$

It should be noted that the expected sums if the dice are fair, are determined from the distribution of *x* as in the table below:

Sum (x)	***P(x)***
2	$^{1}/_{36}$
3	$^{2}/_{36}$
4	$^{3}/_{36}$
5	$^{4}/_{36}$
6	$^{5}/_{36}$
7	$^{6}/_{36}$
8	$^{5}/_{36}$
9	$^{4}/_{36}$
10	$^{3}/_{36}$
11	$^{2}/_{36}$
12	$^{1}/_{36}$

To obtain the expected frequencies, the *P(x)* is multiplied by the total number of trials

Sum (x)	**Observed frequency (O)**	***P(x)***	***Expected Frequency (P(x).500)***
2	15	$^{1}/_{36}$	13.9
3	35	$^{2}/_{36}$	27.8
4	49	$^{3}/_{36}$	41.7
5	58	$^{4}/_{36}$	55.6
6	65	$^{5}/_{36}$	69.5
7	76	$^{6}/_{36}$	83.4
8	72	$^{5}/_{36}$	69.5
9	60	$^{4}/_{36}$	55.6
10	35	$^{3}/_{36}$	41.7
11	29	$^{2}/_{36}$	27.8
12	6	$^{1}/_{36}$	13.9

Recall that $\chi_i^2 = (O_i - E_i)^2/E_i$

Therefore $\chi_1^2 = (O_1 - E_1)^2/E_1 = (15 - 13.9)^2/13.9 = 0.09$

$\chi_2^2 = (O_2 - E_2)^2/E_2 = (35 - 27.8)^2/27.8 = 1.86$

$\chi_3^2 = (O_3 - E_3)^2/E_3 = (49 - 41.7)^2/41.7 = 1.28$

$\chi_4^2 = (O_4 - E_4)^2/E_4 = (58 - 55.6)^2/55.6 = 0.10$

$$\chi_5^2 = (O_5 - E_5)^2/E_5 = (65 - 69.5)^2/69.5 = 0.29$$

$$\chi_6^2 = (O_6 - E_6)^2/E_6 = (76 - 83.4)^2/83.4 = 0.66$$

$$\chi_7^2 = (O_7 - E_7)^2/E_7 = (72 - 69.5)^2/69.5 = 0.09$$

$$\chi_8^2 = (O_8 - E_8)^2/E_8 = (60 - 55.6)^2/55.6 = 0.35$$

$$\chi_9^2 = (O_9 - E_9)^2/E_9 = (35 - 41.7)^2/41.7 = 1.08$$

$$\chi_{10}^2 = (O_{10} - E_{10})^2/E_{10} = (29 - 27.8)^2/27.8 = 0.05$$

$$\chi_{11}^2 = (O_{11} - E_{11})^2/E_{11} = (6 - 13.9)^2/13.9 = 4.49$$

To calculate the overall Chi-squared value, recall that $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ i.e. we add the individual χ^2 value.

Therefore, $\chi^2 = 0.09 + 1.86 + 1.28 + 0.10 + 0.29 + 0.66 + 0.09 + 0.35 + 1.08 + 0.05 + 4.49$

$\chi^2 = 10.34$

For the critical value, since n=11, *d.f.* = 10

Therefore, table value = 18.3

Decision: since the calculated value which is 10.34 is less than table (critical) value the null hypothesis is accepted.

Conclusion: There is no significant difference between observed and expected frequencies. The slight observed differences occurred due to chance.

Exercise: The following figures show the distribution of digits in numbers chosen at random from a telephone directory:

Digit	0	1	2	3	4	5	6	7	8	9	Total
Frequency	1,02	1,107	997	966	1,075	933	1,107	972	964	853	10,000

Test whether the digits may be taken to occur equally frequently in the directory. The table value of χ^2 for d.f at 5% level of significance is 16.92.

Hint: Set up the null hypothesis that the digits 0, 1, 2, 3, ...9 in the numbers in the telephone directory are uniformly distributed, i.e all digits occur equally frequently in the directory. Then, under the null hypothesis, the expected frequency for each of the digits 0, 1, 2, 3,.....9 is 10,000/10 = 1,000

1.4 Chi-Square test for independence of attributes

Consider a given population consisting of *N* items divided into *r* mutually disjoint (exclusive) and exhaustive classes A_1, A_2, ...*Ar* with respect to (*w.r.t*) the attribute A, so that randomly selected item belongs to one and only one of the attributes A_1, A_2, ...*Ar*. Similarly, let us suppose that the same population is divided into *s* mutually disjoint and exhaustive classes B_1, B_2, ...B_s*w.r.t* another attribute B_s so that an item selected at random possesses one and only one of the attributes B_1, B_2, ...B_s can be represented in the following *r x s* manifold contingency e.g like below:

B A	B_1	B_2	.....	B_j	.....	B_s	*Total*
A_1	$(A_1 B_1)$	$(A_1 B_2)$		(A_1B_j)	.....	(A_1B_s)	(A_1)
A_2	$(A_2 B_1)$	$(A_2 B_2)$	.....	(A_2B_j)	.....	(A_2B_s)	(A_2)
:	:	:		.....	.....	:	:
A_i	$(A_i B_1)$	$(A_i B_2)$	.....	(A_iB_j)	.....	(A_iB_s)	(A_i)
:	:	:		.....	.....	:	:
A_r	$(A_r B_1)$	$(A_r B_2)$	.....	A_rB_j	.....	(A_rB_s)	(A_r)

Total	(B_1)	(B_2)	.....	(B_j)	.....	(B_s)	$\sum_{i=1}^{r}(A_i) = \sum_{j=1}^{s}(B_j) = N$

Where (A_i) is the frequency of the *ith* attribute A_i,i.e, it is, number of persons possessing the attribute A_i , i=1,2, ...r; (B_j) is the number of persons possessing the attribute B_j, j=1,2.....s; and (A_iB_j) is the number of persons possessing both the attributes A_i and B_j ; (i: 1, 2, ...r; j: 1, 2, ..., s)

Under the hypothesis that the two attributes A and B are independent, the expected frequency for (A_i, B_i) is given by

$$E[(A_iB_j)] = N.P\,[A_iB_j] = N.P[A_i \cap B_j] = N.P\,[A_i].\,P[B_j]$$

[By compound probability theorem, since attributes are independent]

$$= \mathrm{N}\, X \frac{(A_i)}{N}\; X\; \frac{(B_i)}{N} = \frac{(A_i)(B_i)}{N}$$

If $(A_iB_j)_o$ denotes the expected frequency of (A_iB_j) then

$$(A_iB_j)_o = \frac{(A_i)(B_i)}{N};\ (i = 1, 2, ...,r;\ j=1,2, ...s)$$

Thus, under the null hypothesis of independence of attributes, the expected frequencies for each of the cell frequencies of the above table can be obtained on using this last equation. The rule in the last can be stated in the words as follows:

"*Under the hypothesis of independence of attributes the expected frequency for any of the cell frequencies can be obtained by multiplying the row totals and the column totals in which the frequency occurs and dividing the product by the total frequency N".*

Here, we have a set of r x s observed frequencies (A_iB_j) and the corresponding expected frequencies $(A_iB_j)_o$. Applying χ^2–test of goodness of fit , the statistic

$$\chi^2 = \sum_i \sum_j \left[\frac{[(A_iB_i) - (AiBj)o]^2}{(AiBj)o}\right]$$

follows χ^2-distribution with (r-1)X(s-1) degrees of freedom.

Comparing this calculated value of χ^2 with the tabulated value for (r-1)X(s-1) *d.f.*and at certain level of significance, we reject or retain the null hypothesis of independence of attributes at that level of significance.

Note: For the contingency table data, the null hypothesis is always set up that the attributes under consideration are independent. It is only under this hypothesis that formula $(A_iB_j)_o = \frac{(A_i)(B_i)}{N}$; *(i = 1, 2, ...,r; j=1,2, ...s)* can be used for computing expected frequencies.

Example: A movie producer is bringing out a new movie. In order to map out her advertising, she wants to determine whether the movie will appeal most to a particular age group or whether it will appeal equally to all age groups. The producer takes a random sample from persons attending a pre-reviewing show of the new movie and obtained the result in the table below. Use Chi-square (χ^2) test to arrive at the conclusion (α=0.05).

	Age-groups (in years)				
Persons	***Under 20***	***20-39***	***40– 59***	***60& over***	***Total***
Liked the movie	320	80	110	200	**710**
Dislikedthe movie	50	15	70	60	**195**
Indifferent	30	5	20	40	**95**
Total	**400**	**100**	**200**	**300**	**1,000**

Solution:

It should be noted that the two attributes being considered here are the age groups of the people and their level of likeness of the new movie. Our concern here is to determine whether the two attributes are independent or not.

Null hypothesis (H_o): Likeness of the of the movie is independent of age group (i.e. the movie appeals the same way to different age group)

Alternative hypothesis (H_a): Likeness of the of the movie depends on age group (i.e. the movie appeals differently across age group)

As earlier explained, to calculate the expected value in the cell of row 1 column 1, we divide the product of row 1 total and column 1 total by the grand total (N) i.e.

$E_{ij} = (A_iB_j)/N$

Therefore, $E_{11} = \frac{710\ X\ 400}{1000} = 284$

$E_{12} = \frac{710\ X\ 100}{1000} = 71$

$E_{13} = \frac{710\ X\ 200}{1000} = 142$

$E_{14} = \frac{710\ X\ 300}{1000} = 213$

$E_{21} = \frac{195\ X\ 400}{1000} = 78$

$E_{22} = \frac{195\ X\ 100}{1000} = 19.5$

$E_{23} = \frac{195\ X\ 200}{1000} = 39$

$E_{24} = \frac{195\ X\ 300}{1000} = 58.5$

$E_{31} = \frac{95\ X\ 400}{1000} = 38$

$E_{32} = \frac{95\ X100}{1000} = 9.5$

$E_{33} = \frac{95\ X200}{1000} = 19$

$E_{34} = \frac{95\ X300}{1000} = 28.5$

We can get a table of expected values from the above computations

Table of expected values

	Under 20	**20-39**	**40-59**	**60 &above**
Like	284	71	142	213
Dislike	78	19.5	39	58.5
Indifferent	38	9.5	19	28.5

χ^2 value = $\sum_i \sum_j [\frac{[(A_i B_i)-(AiBj)o]^2}{(AiBj)o}]$ = χ^2 = $\sum_i \sum_j \frac{(Oij-Eij)^2}{Eij}$ where O_{ij} are the observed frequencies while theE_{ij} are the expected values.

$$\chi_{11}^2 = \frac{(320-284)^2}{284} = 4.56$$

$$\chi_{12}^2 = \frac{(80-71)^2}{71} = 1.14$$

$$\chi_{13}^2 = \frac{(110-142)^2}{142} = 7.21$$

$$\chi_{14}^2 = \frac{(200-213)^2}{213} = 0.79$$

$$\chi_{21}^2 = \frac{(50-78)^2}{78} = 10.05$$

$$\chi_{22}^2 = \frac{(15-19.5)^2}{19.5} = 1.04$$

$$\chi^2_{23} = \frac{(70-39)^2}{39} = 24.64$$

$$\chi^2_{24} = \frac{(60-58.5)^2}{58.5} = 0.04$$

$$\chi^2_{31} = \frac{(30-38)^2}{38} = 1.68$$

$$\chi^2_{32} = \frac{(5-9.5)^2}{9.5} = 2.13$$

$$\chi^2_{33} = \frac{(20-19)^2}{19} = 0.05$$

$$\chi^2_{34} = \frac{(40-28.5)^2}{28.5} = 4.64$$

$\chi^2_{calculated} = \sum_i \sum_j \frac{(Oij-Eij)^2}{Eij}$ =4.56+1.14+7.12+0.79+10.05+1.04+24.64+0.04+1.68+2.13+0.05 +4.64 = **57.97**

Recall, that the *d.f.* is (number of row minus one) X (number of column minus one)

$$\chi^2_{(r-1)(s-1)} = 12.59 \quad \text{(critical value)}$$

Decision: Since the calculated χ^2 value is greater than the table (critical value) we shall reject the null hypothesis and accept the alternative.

Conclusion: It can be concluded that the movie appealed differently to different age groups (i.e. likeness of the movie is dependent on age).

4.0 CONCLUSION

In conclusion, chi-squared analysis has very wide applications which include test of independence of attributes; test of goodness fit; test of equality of population proportion and to test if population has a specified variance among others. This powerful statistical tool is useful in business and economic decision making.

5.0 SUMMARY

In this unit, we have examined the concept of chi-square and its scope. We also look at its methodology and applications. It has been emphasized that it is not just an ordinary statistical exercise but a practical tool for solving day-to-day business and economic problems.

6.0 TUTOR-MARKED ASSIGNMENT

1.A sample of students randomly selected from private high schools and sample of students randomly selected from public high schools were given standardized tests with the following results

Test Scores	**0-275**	**276 - 350**	**351 - 425**	**426 - 500**	**Total**
Private School	6	14	17	9	**46**
Public School	30	32	17	3	**86**
Total	**36**	**46**	**34**	**12**	**128**

H_0: The distribution of test scores is the same for private and public high school students at α=0.05

2. A manufacturing company has just introduced a new product into the market. In order to assess consumers' acceptability of the product and make efforts towards improving its quality, a survey was carried out among the three major ethnic groups in Nigeria and the following results were obtained:

	Ethnic groups				
Persons	***Igbo***	***Yoruba***	***Hausa***	***Ijaw***	***Total***

Acceptthe product	48	76	56	70	**250**
Do not Accept	57	44	74	30	**205**
Total	**105**	**120**	**130**	**100**	**455**

Using the above information, does the acceptability of the product depend on the ethnic group of the respondents? (Take α=1%)

7.0 REFERENCES/FURTHER READING

Spiegel, M. R., Stephens L.J., (2008).*Statistics***.(**4th ed.). New York, McGraw Hill press.

Gupta S.C. (2011). *Fundamentals of Statistics***.(**6th Rev.& Enlarged ed.).Mumbai India, Himalayan Publishing House.

Swift L., (1997).*Mathematics and Statistics for Business, Management and Finance.*London UK, Macmillan.

UNIT 4:

ANALYSIS OF VARIANCE (ANOVA)

CONTENTS

1.0 INTRODUCTION

In day-to-day business management and in sciences, instances may arise where we need to compare means. If there are only two means e.g. average recharge card expenditure between male and female students in a faculty of a University, the typical t-test for the difference of two means becomes handy to solve this type of problem. However in real life situation man is always confronted with situation where we need to compare more than two means at the same time. The typical t-test for the difference of two means is not capable of handling this type of problem; otherwise, the obvious method is to compare two means at a time by using the t-test earlier treated. This process is very time consuming, since as few as 4 sample means would require ${}^4C_2 = 6$, different tests to compare 6 possible pairs of sample means. Therefore, there must be a procedure that can compare all means simultaneously. One such procedure is the analysis of variance (ANOVA). For instance, we may be interested in the mean telephone recharge expenditures of various groups of students in the university such as student in the faculty of Science, Arts, Social Sciences, Medicine, and Engineering. We may be interested in testing if the average monthly expenditure of students

in the five faculties are equal or not or whether they are drawn from the same normal population. The answer to this problem is provided by the technique of analysis of variance. It should be noted that the basic purpose of the analysis of variance is to test the homogeneity of several means.

The term Analysis of Variance was introduced by Prof. R.A Fisher in 1920s to deal with problem s in the analysis of agronomical data. Variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes which may be classified as:

(i) Assignable causes and (ii) chance causes

The variation due to assignable causes can be detected and measured whereas the variation due to chances is beyond the control of human and cannot be traced separately.

2.0 OBJECTIVE

The main objective of this unit is to teach students the theories and application of Analysis of Variance (ANOVA). It is hoped that students should after taking this unit be able to apply ANOVA in solving business and economic problem especially as it concern multiple comparison of means

3.0 MAIN CONTENT

3.1 Assumption for ANOVA test

ANOVA test is based on the test statistic F (or variance ratio). For the validity of the F-test in ANOVA, the following assumptions are made:

(i) The observations are independent.
(ii) Parent population from which observation are taken are normal.
(iii) Various treatment and environmental effects are additive in nature.

ANOVA as a tool has different dimensions and complexities. ANOVA can be (a) One-way classification or (b) two-way classification. However, the one-way ANOVA we will deal with in this course material.

Note

(i) ANOVA technique enables us to compare several population means simultaneously and thus results in lot of saving in terms of time and money as compared to several experiments required for comparing two populations means at a time.

(ii) The origin of the ANOVA technique lies in agricultural experiments and as such its language is loaded with such terms as treatments, blocks, plots etc. However, ANOVA technique is so versatile that it finds applications in almost all types of design of experiments in various diverse fields such as industry, education, psychology, business, economics etc.

(iii) It should be clearly understood that ANOVA technique is not designed to test equality of several population variances. Rather, its objective is to test the equality of several population means or the homogeneity of several independent sample means.

(iv) In addition to testing the homogeneity of several sample means, the ANOVA technique is now frequently applied in testing the linearity of the fitted regression line or the significance of the correlation ratio.

3.2 The one-way classification

Assuming n sample observations of random variable X are divided into k classes on the basis of some criterion or factor of classification. Let the $i$thclass consist of $n_i$observations and let:

X_{ij} = jth member of the ith class; {j=1,2,.....n_i; i= 1,2, ...k}

$n = n_1 + n_2 + \ldots\ldots\ldots\ldots + n_k = \sum_{i=1}^{k} n_i$

The n sample observations can be expressed as in the table below:

Class	***Sample observation***	***Total***	***Mean***
1	$X_{11}, X_{12}, \ldots\ldots\ldots\ldots X_{1n}$	T_1	**Mean** X_1
2	$X_{21}, X_{22}, \ldots\ldots\ldots\ldots X_{2n}$	T_2	**Mean** X_2
:	**: : : :**	**:**	**:**
:	**: : : :**	**:**	**:**
I	$X_{i1}, X_{i2}, \ldots\ldots\ldots\ldots X_{in}$	$T_i = \sum_{i=1}^{n} X_{ij}$	**Mean** $X_i = \frac{T_i}{n_i}$
:	**: : : :**	**:**	**:**
:	**: : : :**	**:**	**:**
K	$X_{k1}, X_{k2}, \ldots\ldots\ldots\ldots X_{kn}$	T_k	**Mean** X_k

Such scheme of classification according to a single criterion is called one-way classification and its analysis of variance is known as one-way analysis of variance.

The total variation in the observations X_{ij}can be split into the following two components:

(i) The variation between the classes or the variation due to different bases of classification (commonly known as treatments in pure sciences, medicine and agriculture). This type of variation is due to assignable causes which can be detected and controlled by human endeavour.

(ii) The variation within the classes, i.e. the inherent variation of the random variable within the observations of a class. This type of variation is due to chance causes which are beyond the control of man.

The main objective of the analysis of variance technique is to examine if there is significant difference between the class means in view of the inherent variability within the separate classes.

Steps for testing hypothesis for more than two means (ANOVA): Here, we adopt the rejection region method and the steps are as follows:

Step1: Set up the hypothesis:

Null Hypothesis: Ho: $\mu_1 = \mu_2 = \mu_3 = \ldots\ldots\ldots\ldots = \mu_k$i.e, all means are equal

Alternative hypothesis: H_1 : At least two means are different.

Step 2: Compute the means and standard deviations for each of the by the formular:

$\overline{X_1}$ =

$\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$; $S_i^2 = \frac{1}{n}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$; $(i = 1,2, \ldots\ldots k)$

Also, compute the mean $\bar{X}$ of all the data observations in the k-classes by the formula:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{\sum_i n_i \overline{X_i}}{\sum_i n_i}$$

Step 3: Obtain the Between ClassesSum of Squares (BSS) by the formula:

$$\textbf{BSS} = n_1(\overline{X_1} - \bar{X})^2 + n_2(\overline{X_2} - \bar{X})^2 + \ldots\ldots\ldots\ldots\ldots + n_{3k}(\overline{X_k} - \bar{X})^2$$

Step 4: Obtain the Between Classes Mean Sum of Squares (MBSS)

$$\boldsymbol{MBSS} = \frac{\textit{Between classes Sum of Square}}{\textit{Degrees of freedom}} = \frac{BSS}{k-1}$$

Step 5: Obtain the Within Classes Sum of Squares (WSS) by the formula:

$$\boldsymbol{WSS} = \sum_i \sum_j (X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k} n_i s_i^2 = n_1 s_1^2 + n_2 s_2^2 \dots\dots\dots + n_k s_k^2$$

Step 6: Obtain the Within Classes Mean Sum of Squares (MWSS)

$$\boldsymbol{MBSS} = \frac{\textit{Within classes Sum of Square}}{\textit{Degrees of freedom}} = \frac{WSS}{n-k}$$

Step 7: Obtain the test statistic *F* or Variance Ratio (V.R)

$$F = \frac{\textit{Between classes Mean Sum of Square}}{\textit{Within classes Mean Sum of Square}} = \frac{\textit{Step 4}}{\textit{Step 6}} \sim F(k-1, n-k)$$

Which follows *F*-distribution with (v_1 = k-1, v_2 = n-k)*d.f* (This implies that the degrees of freedom are two in number. The first one is the number of classes (treatment) less one, while the second *d.f* is number of observations less number of classes)

Step 8: Find the critical value of the test statistic *F* for the degree of freedom and at desired level of significance in any standard statistical table.

If computed value of test-statistic F is greater than the critical (tabulated) value, reject (H_o, otherwise H_o may be regarded as true.

Step 9: Write the conclusion in simple language.

Example 1: To test the hypothesis that the average number of days a patient is kept in the three local hospitals A, B and C is the same, a random check on the number of days that seven patients stayed in each hospital reveals the following:

Hospital A:	8	5	9	2	7	8	2
Hospital A:	4	3	8	7	7	1	5
Hospital A:	1	4	9	8	7	2	3

Test the hypothesis at 5 percent level of significance.

Solution: Let X_{1j}, X_{2j}, X_{3j} denote the number of days the jth patient stays in the hospitals A, B and C respectively

Calculations for various Sum of Squares

X_{1j}	X_{2j}	X_{3j}	$(X_{1j} - \bar{X}_1)^2$	$(X_{2j} - \bar{X}_2)^2$	$(X_{3j} - \bar{X}_3)^2$
8	4	1	4.5796	1	14.8996
5	3	4	0.7396	4	0.7396
9	8	9	9.8596	9	17.1396
2	7	8	14.8996	4	9.8596
7	7	7	1.2996	4	4.5796
8	1	2	4.5796	16	8.1796
2	5	3	14.8996	0	3.4596
Total=ΣX_{1j} = T_1 = 41	$\Sigma X_{2j} = T_2$ = 35	$\Sigma X_{3j} = T_3$ = 41	$\sum_{j=1}^{7}(X_{1j} - \bar{X}_1)^2$ =50.8572	$\sum_{j=1}^{7}(X_{2j} - \bar{X}_2)^2$ =38	=58.8572

$$\bar{X}_1 = \frac{\sum X_{1j}}{n_1} = \frac{41}{7} = 5.86\,;$$

$$\bar{X}_2 = \frac{\sum X_{2j}}{n_2} = \frac{35}{7} = 5$$

$$\bar{X}_3 = \frac{\sum X_{3j}}{n_3} = \frac{34}{7} = 4.86$$

$$\bar{X} = \frac{Grand\ Total}{Total\ number\ of\ observation} = \frac{41+35+34}{7+7+7} = \frac{110}{21} = 5.24$$

Within Sample Sum of Square: To find the variation within the sample, we compute the sum of the square of the deviations of the observations in each sample from the mean values of the respective samples (see the table above)

Sum of Squares within Samples =

$$\sum_{j=1}^{7}(X_{1j} - \bar{X}_1)^2 + \sum_{j=1}^{7}(X_{2j} - \bar{X}_2)^2 + \sum_{j=1}^{7}(X_{3j} - \bar{X}_3)^2$$

= 50.8572 + 38 + 58.8572 = 147.7144 ~ 147.71

Between Sample sum of Squares: $\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$

To obtain the variation between samples, we compute the sum of the squares of the deviations of the various sample means from the overall (grand) mean.

$$(\overline{X_1} - \bar{X})^2 = (5.86 - 5.24)^2 = (0.62)^2 = 0.3844;$$

$$(\overline{X_2} - \bar{X})^2 = (5 - 5.24)^2 = (-0.24)^2 = 0.0576;$$

$$(\overline{X_3} - \bar{X})^2 = (4.86 - 5.24)^2 = (-0.38)^2 = 0.1444;$$

Sum of square Between Samples (hospitals):

$$\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 = n_1(\overline{X_1} - \bar{X})^2 + n_2(\overline{X_2} - \bar{X})^2 + n_3(\overline{X_3} - \bar{X})^2$$

= 7(0.3844) + 7(0.0576) + 7(0.1444)

= 2.6908 + 0.4032 + 1.0108 = 4.1048 = 4.10

Total Sum of Squares: = $\sum_i \sum_j (X_{ij} - \bar{X}_i)^2$

The total variation in the sample data is obtained on calculating the sum of the squares of the deviations of each sample observation from the grand mean, for all the samples as in the table below:

X_{1j}	$(X_{1j} - \bar{X})^2$ $=(X_{1j} - 5.24)^2$	X_{2j}	$(X_{2j} - \bar{X})^2$ $=(X_{2j} - 5.24)^2$	X_{3j}	$(X_{3j} - \bar{X})^2$ $=(X_{3j} - 5.24)^2$
8	7.6176	4	1.5376	1	17.9776
5	0.0576	3	5.0176	4	1.5376
9	14.1376	8	7.6176	9	14.1376
2	10.4976	7	3.0976	8	7.6176
7	3.0976	7	3.0976	7	3.0976
8	7.6176	1	17.9776	2	10.4976
2	10.4976	5	0.0576	3	5.0176
Total = 41	**53.5232**	**35**	**38.4032**	**34**	**59.8832**

Total sum of squares (TSS) =

$$\sum (X_{1j} - \bar{X})^2 + \sum (X_{2j} - \bar{X})^2 + \sum (X_{3j} - \bar{X})^2$$

= 53.5232 + 38.4032 + 59.8832 = 151.81

Note: Sum of Squares Within Samples + S.S Between Samples = 147.71 + 4.10 =151.81

= Total Sum of Squares

Ordinarily, there is no need to find the sum of squares within the samples (i.e, the error sum of squares), the calculations of which are quite tedious and time consuming. In practice, we find the total sum of squares and between samples sum of squares which are relatively

simple to calculate. Finally within samples sum of squares is obtained by subtracting Between Samples Sum of Squares from the Total Sum of Squares:

W.S.S.S = T.S.S – B.S.S.S

Therefore, Within Sample (Error) Sum of Square = 151.8096 – 4.1048 = 147.7044

Degrees of freedom for:

Between classes (hospitals) Sum of Squares = k-1 = 3-1=2

Total Sum of Squares = n-1 = 21-1 = 20

Within Classes (or Error) Sum of Squares = n-k = 21 – 3= 18

ANOVA TABLE

Sources of variation(1)	*d.f*(2)	Sum of Squares(S.S) (3)	Mean Sum of Squares(4) = $\frac{(3)}{(2)}$	Variance Ratio(F)
Between Samples (Hospitals)	3-1 =2	4.10	$\frac{4.10}{2} = 2.05$	$\frac{2.05}{8.21} = 0.25$
Within Sample (Error)	20-2=18	147.71	$\frac{147.71}{18} = 8.21$	
Total	21-1=20	151.81		

Critical Value: The tabulated (critical) value of *F* for *d.f* (v_1=2, v_2=18) *d.f* at 5% level of significance is 3.55

Since the calculated *F* = 0.25 is less than the critical value 3.55, it is not significant. Hence we fail to accept Ho.

However, in cases like this when MSS between classes is less than the MSS within classes, we need not calculate *F* and we may conclude that the means $\overline{X_1}$, $\overline{X_2}$ and $\overline{X_3}$ do not differ significantly. Hence, H_o may be regarded as true.

Conclusion: $H_o : \mu_1 = \mu_2 = \mu_3$, may be regarded as true and we may conclude that there is no significant difference in the average stay at each of the three hospitals.

Critical Difference: If the classes (called treatments in pure sciences) show significant effect then we would be interested to find out which pair(s) of treatment differ significantly. Instead of calculating Student's *t* for different pairs of classes (treatments) means, we calculate the Least Significant Difference (LSD) at the given level of significance. This LSD is also known as Critical Difference (CD).

The LSD between any two classes (treatments) means, say $\overline{X_i}$ and $\overline{X_j}$ at level of significance 'α' is given by:

LSD ($\overline{X_i}$ -$\overline{X_j}$) = [The critical value of t at level of significance α and error d.f] X [S.E ($\overline{X_i}$ - $\overline{X_j}$)]

Note: S.E means Standard Error. Therefore, the S.E ($\overline{X_i}$ -$\overline{X_j}$) above mean the standard error of the difference between the two means being considered.

$$= t_{n-k\,(\alpha/2)} \times \sqrt{MSSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

MSSE means sum of squares due to Error

If the difference $\left|\overline{X_i} - \overline{X_j}\right|$ between any two classes (treatments) means is greater than the LSD or CD, it is said to be significant.

Another Method for the computation of various sums of squares

Step 1: Compute: $G = \sum_i \sum_j X_{ij} = Grand\ Total\ of\ all\ observations$

Step 2: Compute Correction Factor (CF) = $\frac{G^2}{n}$, where n = $n_1+n_2=.....n_k$, is the total number of observations.

Step 3: Compute Raw Sum of Square (RSS) = $\sum_i \sum_j X_{ij}{}^2$ = Sum of squares of all observations

Step 4: Total Sum of Square = $\sum_i \sum_j (X_{ij} - \overline{X})^2 = RSS - CF$

Step 5: Compute

$T_i = \sum_{j=1}^{n_i} X_{ij} = The\ sum\ of\ all\ observations\ in\ the\ ith\ class; (i = 1,2,..k)$

Step 6: Between Classes (or Treatment) Sum of Squares = $\sum_{i=1}^{k} \frac{T_i^2}{n_i} - CF$

$$= \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \cdots \ldots\ldots\ldots\ldots \frac{T_k^2}{n_k} - CF$$

Step 7: Within Classes or Error Sum of Squares = Total S.S – Between Classes S.S

The calculations here are much simpler and shorter than in the first method

Application: Let us now apply this alternative method to solve the same problem treated earlier.

n= Total number of observation = 7 + 7 + 7 = 21

Grand Total (*G)* =

$$\sum_i \sum_j X_{ij} = (8+5+9+2+7+8+2) + (4+3+8+7+7+1+5) + (1+4+9+8+7+2+3 = 110$$

Correction Factor = (CF) = $\frac{G^2}{n} = \frac{110^2}{21} = 576.1905$

Raw Sum of Square (RSS) = $\sum_i \sum_j X_{ij}{}^2$

$$= (8^2 + 5^2 + 9^2 + 2^2 + 7^2 + 8^2 + 2^2) + (4^2 + 3^2 + 8^2 + 7^2 + 7^2 + 1^2 + 5^2)$$

$$+ (1^2 + 4^2 + 9^2 + 8^2 + 7^2 + 2^2 + 3^2)$$

$$= 291 + 213 + 224 = 728$$

Total Sum of Square (TSS) = RSS – CF = 728 – 576.1905 = 151.8095

Between Classes (hospitals) Sum of Squares = $\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} - CF$

But $T_1 = \sum_j X_{1j} = 41,\ T_2 = \sum_j X_{2j} = 35, T_3 = \sum_j X_{3j} = 34,$

Therefore, BCSS $= \frac{41^2}{7} + \frac{35^2}{7} + \frac{34^2}{7} - CF$

$= \frac{1681+1225+1156}{7} - 576.1905 = 580.2857 - 576.1905 = .0952$

Therefore, Within Classes (hospitals) Sum of Squares or Error S.S = TSS – BCSS

$$= 151.8095 – 4.0957 = 147.7138$$

Having arrived at the same Sums of Squares figures, computations can proceed as done earlier.

Example 2: The table below gives the retail prices of a commodity in some shops selected at random in four cities of Lagos, Calabar, Kano and Abuja. Carry out the Analysis of Variance (ANOVA) to test the significance of the differences between the mean prices of the commodity in the four cities.

City	***Price per unit of the commodity in different shops***			
Lagos	9	7	10	8
Calabar	5	4	5	6
Kano	10	8	9	9
Abuja	7	8	9	8

If significant difference is established, calculate the Least Significant Difference (LSD) and use it to compare all the possible combinations of two means (α=0.05).

Solution:

Using the alternative method of obtaining the sum of square

City	***Price per unit of the commodity in different shops***				***Total***	***Means***
Lagos	9	7	10	8	34	8.5
Calabar	5	4	5	6	20	5
Kano	10	8	9	9	36	9
Abuja	7	8	9	8	32	8

Grand Total (G) = $\sum_i \sum_j X_{ij}$ = (9+7+10+8) + (5+4+5+6) + (10+8+9+9) + (7+8+9+8)

= 34 + 20 + 36 +32

= 122

Correction Factor (CF) = $\frac{G^2}{n}$

= $\frac{(122)^2}{16}$

= $\frac{14,884}{16}$

= 930.25

Raw Sum of Square (RSS) = $\sum_i \sum_j X_{ij}{}^2$

= $(9^2 + 7^2 + 10^2 + 8^2) + (5^2 + 4^2 + 5^2 + 6^2) + (10^2 + 8^2 + 9^2 + 9^2) +$

$(7^2 + 8^2 + 9^2 + 8^2)$

= 294 + 102 + 326 + 258

RSS = 980

Total Sum of Square (TSS) = RSS – CF

= 980 – 930.5

TSS= 49.75

Between Classes (cities) Sum of Squares = $\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} - CF$

$= \frac{34^2}{4} + \frac{20^2}{4} + \frac{36^2}{4} + \frac{32^2}{4} - CF$

$= \frac{1156}{4} + \frac{400}{4} + \frac{1296}{4} + \frac{1024}{4} - CF$

BCSS = 289 + 100 + 324 + 256 – 930.25

= 969 – 930.25

BCSS = 38.75

Within Class (cities) or Error Sum of Squares = TSS – BCSS

= TSS –BCSS

= 49.75 – 38.75

WSS = 11

Between Class Mean Sum of Square Error = $\frac{BCSS}{k-1}$; where k is the number of classes

$= \frac{38.75}{4-1} = \frac{38.75}{3}$

= 12.92

Within Class Mean Sum of Square Error (WCMSSE) = $\frac{WSS}{n-k} = \frac{11}{16-4}$

= 0.92

Variance Ratio ($F_{calculated}$) = $\frac{BCMSSE}{WCMSSE}$

$$F_{calculated} = \frac{12.92}{0.92}$$

$$F_{calculated} = 14.04$$

F-table (critical value) = $F_{(v1, v2, \alpha)} = F_{(3, 12, 0.05)} = 3.49$

Decision: Sincethe computed *F* is greater than the table value $F_{(v1, v2, \alpha)}$ **,** the null hypothesis is rejected and the alternative is accepted.

Conclusion: At least one of the means is significantly different from others.

$$\textbf{LSD} = t_{n-k(\alpha/2)}.\ S.E(\bar{X}_i - \bar{X}_j)$$

But the standard error of $(\bar{X}_i - \bar{X}_j) = \sqrt{WCMMSE\ X\ \frac{1}{n_i} + \frac{1}{n_j}}$

Therefore, **LSD** = 2.18 X $\sqrt{0.92\ X\ \frac{1}{4} + \frac{1}{4}}$

= 2.18 X $\sqrt{0.46}$

= 2.18 X 0.678

LSD = 1.48

Comparison between different means

Cities	**Absolute Difference**	**Comparison**	**Conclusion**
Lagos and Calabar	$\|8.5 - 5\| = 3.5$	> LSD	Significant

Lagos and Kano	$\|8.5 - 9\| = 0.5$	< LSD	Not Significant
Lagos and Abuja	$\|8.5 - 8\| = 0.5$	< LSD	Not Significant
Calabar and Kano	$\|5 - 9\| = 4$	> LSD	Significant
Calabar and Abuja	$\|5 - 8\| = 3$	> LSD	Significant
Kano and Abuja	$\|9 - 8\| = 1$	< LSD	Not Significant

4.0 CONCLUSION

The unit has espoused the theory and application of Analysis of Variance in statistics with special emphasis on its application in the comparison of more than two means.

5.0 SUMMARY

In summary, ANOVA is very useful in the multiple comparison of mean among other important uses in both social and applied sciences.

6.0 TUTOR-MARKED ASSIGNMENT

Concord Bus Company just bought four different Brands of tyres and wishes to determine if the average lives of the brands of tyres are the same or otherwise in order to make an important management decision. The Company uses all the brands of tyres on randomly selected buses. The table below shows the lives (in '000Km) of the tyres:

Brand 1: 10, 12, 9, 9

Brand 2: 9, 8, 11, 8, 10

Brand 3: 11, 10, 10, 8, 7

Brand 4: 8, 9, 13, 9

Test the hypothesis that the average life for each of brand of tyres is the same. Take $\alpha = 0.01$

7.0 REFERENCES / FURTHER READINGS

Spiegel, M. R., Stephens L.J., (2008).*Statistics*.**(**4th ed.). New York, McGraw Hill press.

Gupta S.C. (2011). *Fundamentals of Statistics*.**(**6th Rev.& Enlarged ed.).Mumbai India, Himalayan Publishing House.

Swift L., (1997).*Mathematics and Statistics for Business, Management and Finance.*London UK, Macmillan.