NATIONAL OPEN UNIVERSITY OF NIGERIA

COURSE CODE: AEA 501

COURSE TITLE: STATISTICS FOR SOCIAL SCIENCES

COURSE GUIDE

AEA 501
STATISTICS FOR SOCIAL SCIENCES

NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
University Village
Plot 91, Cadastral Zone,
Nnamdi Azikwe Express way
Jabi-Abuja

Lagos Office
14/16 Ahmadu Bello Way
Victoria Island, Lagos

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

Published by
National Open University of Nigeria

Printed 2016

ISBN: 978-058-409-9

CONTENTS **PAGE**

INTRODUCTION

AEA 501: Statistics for Social Sciences is a three (3) credit unit course designed for 500 level students in the school of Agricultural Sciences. The course will expose you to an understanding of many concepts in Statistics for Social Sciences and their applications.

The course consists of two major parts, i.e. The Course Guide and The Study Guide. The Study Guide consists of twelve modules and thirty units.

This course guide tells you briefly what the course is all about, what course materials you will be using and how you can work your way through these materials with minimum assistance. It suggests some general guidelines for the amount of time you might spend in order to successfully complete each unit of the course. It also gives you some guidance on your Tutor Marked Assignment (TMA).

WHAT YOU WILL LEARN IN THIS COURSE

AEA 501 Statistics for Social Sciences consist of twelve major components arranged in modules:

The first module which is on the background, roles, scope and limitations of statistics will introduce you to the definition and general background of statistics.

The approach to data collection, classification and presentation, forms the major part of the study in the second module. You shall lean about methods of data collection and presentation of data.

The third module of this course, which is on measures of central tendency, will discuss topics such as measures of central tendency, the arithmetic mean, weighted arithmetic mean, geometric mean and the harmonic mean among others.

The fourth module of the course will focus on measures of dispersion and its application. You will learn how to calculate the range from a frequency distribution and also acquire skills in the calculation of mean deviation.

Module five focuses on population, sample and sampling where you will learn what population or universe is, learn how to sample from a population, and understand the reasons for sampling.

Module six is about probability. This module will introduce you to some concepts in probability and, the different types of probability.

The seventh module will discuss factorial, permutation, combination and mathematical expectations. Here, you will be provided basic explanations and calculations of permutation using formula and how to apply combinations formula for practical applications.

Module eight is on binomial, poison and normal distributions. This module will make you aware of some definitions and application of Poisson distribution. You will also learn about normal distribution and its properties.

In module nine, you shall learn about central limit theory, confidence interval and hypothesis testing. From this module, you will acquire skills in statistical estimations, learn basic steps involved in hypothesis testing and how to calculate confidence interval.

Module 10 is about student's t-test, z distribution and f distribution. In this module, you shall learn how to test hypothesis using student's t test, learn how to apply Z distribution and acquire skills in the calculation of F distribution.

Module 11 introduces you to chi square analysis and its application. This module also touches the basic aspect of chi square which is called test of goodness of fit.

The last module which is twelve will focus on correlation and regression analysis. This module will help you understand the concept surrounding correlation and regression analysis and its application in calculations and how to interpret the regression result.

COURSE AIMS

The aim of this course is to give a clear understanding of statistics for social sciences. This aim will be achieved through:

- understanding the background, roles, scope and limitations of statistics
- understanding the data collection, classification and presentation
- understanding the measures of central tendency
- understanding the measures of dispersion
- understanding population, sample and sampling techniques
- understanding probability
- understanding factorial, permutation, combination and mathematical expectations
- understanding binomial, poisson and normal distributions
- understanding the central limit theory, confidence interval and hypothesis testing
- understanding the student's T-Test, Z distribution and F test distribution
- understanding chi square analysis
- understanding correlation and regression analysis

OBJECTIVES

In order to achieve the aims of this course, there are sets of overall objectives. The unit objectives are always included in the beginning of the unit. You need to read them before you start working through the unit. You may also need to refer to them during your study of the unit to check your progress. You should always look at the unit objectives after completing a unit. In doing so, you will be sure that you have followed the instruction in

the unit. And by meeting these objectives you should have achieved the aims of the course as a whole. However, the general objectives of the course include:

- define statistics and explain the roles of statistics
- identify the various areas of application of statistics
- explain vividly, the different methods of data collection and presentation
- differentiate between arithmetic, weighted, geometric and harmonic means
- use cumulative frequency polygon (ogive) to determine the median of the distribution
- differentiate between quartile and percentile
- state the relationship between mean, median and mode
- Differentiate one sampling techniques from the other
- Explain the different types of probability
- Outline the conditions under which Poisson distribution is applied
- Define what statistical hypothesis is all about
- Differentiate between type I and type II errors
- Outline the basic steps involved in hypothesis testing
- Outline the conditions under which Z distribution is applied
- Differentiate Z distribution from student's t-distribution
- Define what correlation analysis is all about
- Interpret the strength and direction of relationship of the correlation
- Define what regression analysis is all about
- Explain how regression analysis is applied in statistics
- Describe briefly the direction, strength and rate of change of one variable in relation to another.
- Determine the regression coefficient b

COURSE REQUIREMENTS

To complete this course you are required to read the study units, read suggested books and other materials that will help you achieve the stated objectives. Each unit contains self assessment exercises and Tutor Marked Assignment (TMA) which you are encouraged to answer and at intervals as you progress in the course, you are required to submit assignment for assessment purpose. There will be a final examination at the end of the course.

COURSE MATERIAL

You will be provided with following materials for this course:

1. **Course Guide**
   The material you are reading now is called course guide, which introduce you to this course
2. **Study Guide**

The textbook prepared for this course by National Open university of Nigeria is called Study Guide. You will be given a copy of the book for your personal use.

3. **Text Books**
At the end of each unit, there is a list of recommended textbooks which though not compulsory for you to acquire or read, are necessary as supplements to the course materials.

STUDY UNITS

There are thirty (30) study units in this course divided into twelve modules as follows:

Module 1 Background, Roles, Scope and Limitations of Statistics

Unit 1 General Background and Roles of Statistics
Unit 2 Scope of Statistics and its limitations

Module 2 Data Collection, Classification and Presentation

Unit 1 Techniques of data collection
Unit 2 Data Presentation and tabulation

Module 3 Measures of Central Tendency

Unit 1 Measures of Central Tendency (The Arithmetic, Weighted, Geometric and Harmonic Mean)
Unit 2 Measures of Central Tendency (Median)
Unit 3 Quartile and Percentile
Unit 4 The Mode and Relationship Between Mean, Median and Mode

Module 4 Measures of Dispersion

Unit 1 Measure of Dispersion (the Range and Mean Deviation)
Unit 2 Standard Deviation and Variance as a Measure of Dispersion

Module 5 Population, Sample and Sampling

Unit 1 Population and Sample
Unit 2 Types of Sampling

Module 6 Probability

Unit 1 Probability and Types of Probability
Unit 2 Rules of Probability

Module 7 Factorial, Permutation, Combination and Mathematical Expectations

Unit 1 Factorial and Permutations
Unit 2 Combinations and Mathematical Expectations

Module 8 Binomial, Poisson and Normal Distributions

Unit 1 Binomial Distribution
Unit 2 Poisson Distribution
Unit 3 Normal Distribution

Module 9 Central Limit Theory, Confidence Interval and Hypothesis Testing

Unit 1 Central Limit Theory
Unit 2 Statistical Estimation and Confidence Interval
Unit 3 Test of Hypothesis

Module 10 Student's T-Test, Z Distribution and F Distribution

Unit 1 Student's t distribution
Unit 2 Z distribution
Unit 3 F distribution

Module 11 Chi Square Analysis

Unit 1 Test of goodness of fit (Chi square test)
Unit 2 Test of independence (Chi square test)

Module 12 Correlation and Regression Analysis

Unit 1 Correlation Analysis
Unit 2 Regression Analysis
Unit 3 Regression Equation

At intervals in each unit, you will be provided with a number of exercises or self-assessment questions. These are to help you test yourself on the materials you have just covered or to apply it in some way. The value of these self-test is to help you evaluate your progress and to re-enforce your understanding of the material. At least one tutor-marked assignment will be provided at the end of each unit. The exercise and the tutor-marked assignment will help you in achieving the stated objectives of the individual unit and that of the entire course.

TEXTBOOKS AND REFERENCES

For detailed information about the areas covered in this course, you are advised to consult more recent edition of the following recommended books:

Afonja B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Evans Brothers Nigeria Publishing Limited, Ibadan, Nigeria, 317 pp.

Asika, N. (2006). *Research Methodology in the Behavioural Sciences*, Longman Nigeria Plc, Lagos. 194 pp

Ayandike, R. N. C. (2009). *Statistical Methods for Social and Environmental Sciences.* Spectrum Books Limited, Ibadan, 390 pp.

Gupta, C. B. And Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). VIKAS Publishing House PVT Limited, New Delhi, pp829.

Spiegel, M. R. And Stephens, L. J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* McGraw.Hill, USA. 538 pp

ASSESSMENT

There are two components of assessment for this course.

1. Tutor-Marked Assignment (TMA)
2. End of Course Examination

TUTOR MARKED ASSIGNMENT (TMA)

The TMA is the continuous assignment component of this course. It account for 30 percent of the total score. You will be given about four TMAs to answer where the facilitator will pick the best three for you. You must submit all your TMAs before you are allowed to sit for the end of course examination.

FINAL EXAMINATION AND GRADING

This examination concludes the assessment for the course. It constitutes 70 percent of the whole course. The final examination for the course will be 3hours duration and consist of seven theoretical questions and you are expected to answer five questions. The examination will consist of questions, which reflect the Tutor Marked Assignments that you might have previously encountered and other questions within the course covered areas. All areas of the course will be covered by the assignment. You are to use the time between finishing the last unit and sitting for the examination to revise the entire course. You might find it useful to review your Tutor Marked Assignments before the examination. The final examination covers information from all parts of the course.

SUMMARY

AEA 501: Statistics for Social Sciences is designed to provide background understanding of Statistics and its application to the social sciences. By the time you complete studying this course, you will be able to answer basic questions concerning the course in under the following:

1. Background of Statistics
2. Data Collection, Classification and Presentation
3. Measures of Central Tendency and Dispersion
4. Population, Sample and Sampling
5. Probability
6. Factorial, Permutation, Combination and Mathematical Expectations
7. Binomial, Poisson and Normal Distributions
8. Central Limit Theory, Confidence Interval and Hypothesis Testing
9. Student's T-Test, Z Distribution, F Distribution, Chi Square Analysis, Correlation and Regression Analysis

MAIN COURSE

CONTENTS **PAGE**

MODULE 1 BACKGROUND, ROLES, SCOPE AND LIMITATIONS OF STATISTICS

Unit 1 General Background and Roles of Statistics
Unit 2 Scope of Statistics and its Limitations

UNIT 1 GENERAL BACKGROUND AND ROLES OF STATISTICS

CONTENTS

1.0 INTRODUCTION

This unit provides you a general understanding of what this unit is all about and how it fits into the course as a whole. This unit provides the definition and general background of statistics. Comprehending the definition and general background of any topic is very important. It enables you to have a clear understanding of what the course is all about. It is hoped that at the end of the unit you would have achieved the objectives stated below.

2.0 OBJECTIVES

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

- define statistics
- describe major steps in statistical analysis process
- explain vividly the types of statistics
- discuss statistics, its scopes and role in the social sciences fields.

3.0 MAIN CONTENT

3.1 Definition of Statistics

Statistics is concerned with scientific method of collecting, organising, summarising, presenting and analysing masses of numerical data so as to comprehend the essential features and relationship of the data. It can as well be defined as the study of the methods of collecting and analysis of data in such a way as to minimise any uncertainty in the conclusion drawn from the data. A mass of data has little or no meaning until they are subjected to statistical analysis. Major steps in statistical analysis are:

i. data collection
ii. organising and summarising data
iii. analysing and interpreting the result
iv. using the result to make rational decision.

3.2 Types of Statistics

There are two major classifications of statistics: descriptive and inferential statistics.

i. **Descriptive statistics**: This describes and summarises the entire population through tables or charts, in order to bring out the important facts about the data. Examples of descriptive statistics are mean, median, mode and percentages.

ii. **Inferential statistics**: This is sometimes called objective or analytical statistics which is a method for studying only a part of the population in order to draw conclusions on the population based on the analysis of the sample data. Examples of inferential statistics are probability distributions, correlation and regression analysis. The main purpose of inferential statistics is to make an inference about a population based on the information contained in the sample.

3.3 Roles of Statistics

The following are some of the roles of statistics:

i. Statistics simplifies complex mass of data and presents them in a comprehensive way that they are at once made easy to comprehend and interpret. Instead of having a large raw data, the data are prepared in percentages and means which can be grasped

more easily than a mass of data. Also, statistical analyses in the form of histogram, bar chart, or pie chart, make it easier for you to understand.

ii. Statistics presents data in more comprehensive and definite form: Statistics made conclusion that are stated numerically and more convincing than conclusions stated qualitatively. For example, it is more attractive and convincing to say 85% of candidates that sat for WAEC passed in 2009 than saying most candidates that sat for WAEC in 2009 passed.

iii. Statistics interpret conditions that are more presentable: Statistics present conditions in an attractive ways such as pie chart and histogram or bar charts of the phenomenon under investigation. Also, certain conditions are proved statistically to find out the probability of future occurrence of such situation so that necessary actions could be taken to prevent future occurrence of such conditions.

iv. It provides easy way of classifying numerical data: The method of classification in statistics provides the salient features of the variables that is under consideration. For example, statistical methods provide appropriate method of classifying two or more data by bringing out the maximum, minimum and the standard deviation of the various categories.

v. It provides an easy way of comparing data: Some data may be meaningless unless they are subjected to statistical analysis before they can be compared with similar data at other places. Statistics made an easy way of relating two different masses of numerical data by comparing some relevant information from the two data such as comparing their means, median and mode of their distribution.

SELF-ASSESSMENT EXCERCISE

i. Differentiate between descriptive and inferential statistics with relevant examples.
ii. How important is statistics in data analysis.

4.0 CONCLUSION

This unit has introduced you to the meaning and background of statistics. From this discussion, you must have learnt the definition, background, major types and roles of statistics.

5.0 SUMMARY

The main points in this unit are:

1. The term 'statistics' is concerned with scientific method of collecting, organising, summarising, presenting and analysing masses of numerical data.
2. Statistics is broadly classified as descriptive statistics and inferential statistics. The descriptive statistics is used to analyse and summarise data while inferential statistics is used to draw conclusions on the population based on the analysis of the sample.
3. The various roles of statistics include simplification of complex mass of data, presentation of data in more comprehensive and definite form, easy interpretation of data and easy way of comparing data.

6.0 TUTOR-MARKED ASSIGNMENT

1. Mention and explain some major steps in statistical analysis.
2. Explain the two broad classifications of statistics.
3. What are the roles of statistics?

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics*, David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 2 SCOPE OF STATISTICS AND ITS LIMITATIONS

CONTENTS

1.0 INTRODUCTION

This unit will make you aware of the scope of statistics or how wide its area of application is. This is very important because you need to know the areas in which statistics is applicable. The objectives below specify what you are expected to learn after going through this unit.

2.0 OBJECTIVES

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

- identify the various areas of application of statistics
- state the limitations of statistics.

3.0 MAIN CONTENT

3.1 Statistics as a Subject and its Application

Statistics has very wide application in many fields. It is very useful in government's analysis and publication of many countries. For example, in population and housing, wages and salaries, budgets, education, agriculture, health, births and deaths of a country, etc. statistical tools are highly useful in population censuses and sample survey of national and international assignments. The different fields of application of statistics are:

i. **Economics**: Statistical data are highly useful in the understanding of economic policy and economic problems. In economics, numerical data are usually analysed statistically for ease of

understanding. For example, volume of trade in Nigeria, price of commodities in certain year, wages and salaries of workers etc.

ii. **Agricultural and biological sciences**: In Agriculture, statistics is very important in crop experimental designs. It is used in comparing the results of crops from different fields and different designs. Statistical tools are also very useful in Biological Sciences as a means of testing the significance of results of experiments. It is also used in determining the relationship between the growth of different levels of feeding. It is applied in the estimation of fish population in a lake.

iii. **Physical sciences**: Some statistical methods were applied and developed in the field of physical sciences like Physics, Geology, Chemistry, etc. In the recent time, physical sciences are making good use of statistics in complex cases of molecular, nuclear and atomic structures.

3.2 Limitations of Statistics

Despite the usefulness of statistics in many fields, it also has some limitations and cannot be an answer to the whole affairs of the world. The following are some of its limitations:

i. Statistics deals with group or set of data and attach less importance to individual items. Statistics proves inadequate where the knowledge of individual items becomes necessary. It is most suited to those problems where aggregate characteristics are required.

ii. Statistics deals mainly with quantitative or numerical data: It is not all subjects that can be expressed numerically; there are situations where qualitative data are required. For example, poverty, intelligence and health are all qualitative data which cannot be directly quantified. So, these types of data are not suitable for statistical analysis.

iii. Error during sampling could be to establish wrong conclusions if not handled by experts. Incorrect application of methods could lead to the drawing of wrong conclusions in statistics.

iv. Statistical data in most cases is usually approximated and not very exact. More emphases is usually laid on sampling method of data collection, that means that if a limited number of items are selected, the result of the sample may not be a true representation of the population.

SELF-ASSESSMENT EXERCISE

Statistics as a subject has a wide field of application, discuss this with relevant examples.

4.0 CONCLUSION

In this unit, you have learnt the various areas of application of statistics. Also, we learned the limitations of statistics despite its usefulness. From these discussions, you will now be able to tell us the areas of application of statistics as well as the limitations.

5.0 SUMMARY

A summary of the main points in this unit are:

- Statistics is applicable to various fields such as Economics, Agricultural and Biological sciences as well as physical sciences.
- The limitations of statistics include dealing with group of data with less attention to individual observation, it deals with quantitative data which are not applicable to all subjects and sampling error can lead to wrong conclusion in statistics.

6.0 TUTOR-MARKED ASSIGNMENT

1. Explain the different fields of application of statistics.
2. What are the limitations of statistics?

7.0 REFERENCES/FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics*, David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

MODULE 2 DATA COLLECTION, CLASSIFICATION AND PRESENTATION

Unit 1 Techniques of Data Collection and Data Classification
Unit 2 Data Presentation and Tabulation

UNIT 1 TECHNIQUES OF DATA COLLECTION AND DATA CLASSIFICATION

CONTENTS

1.0 INTRODUCTION

We have studied the general background, scope and limitations of statistics in module 1. In module 2 we shall learn about methods of data collection and presentation of data. Unit 1 discusses in detail the definition and classification of data. It is expected that at the end of this unit, you should have achieved the stated objectives of this unit.

2.0 OBJECTIVES

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

- differentiate between raw data and information
- explain vividly, the different methods of data collection
- outline briefly the different methods of classifying data.

3.0 MAIN CONTENT

3.1 Techniques of Data Collection

Raw data: Raw data are unprocessed bits of information about some topic or event. The raw data need to be processed by subjecting it to some statistical analysis before it can make meaningful interpretation. A piece of data is called a score or observation while a collection of pieces of data pertaining to a topic is called variable. For example, the age of an individual is a score or observation, while the age of the entire group is a variable. Data can be collected from respondents through the use of structured questionnaires, group discussion (interview method), through phone or through email.

3.2 Classification of Data

Data could be classified as **quantitative** and **qualitative**. A quantitative data can be counted or measured e.g. age, income, years, farm size, inflation rate etc. Qualitative data on the other hand cannot be counted or measured but can only be explained, male or female, scale of operation, opinion of group of people etc. Due to the fact that statistics deals with numerical data, qualitative data are usually converted into quantitative data by coding it or assigning value to it. For example, male = 1 and female = 0.

Data could also be classified as **primary** and **secondary** data. Primary data are first-hand information. They are data collected by investigator himself or ask someone to collect for him, which have not been documented or analysed by someone else before. Primary data can be obtained through observation, interview method, questionnaire method, participatory rural appraisal method, etc. Secondary data on the other hand is second hand information. It is an existing data already collected by someone else. An existing data may be in published or unpublished form. For example, in a study of the ages of graduating students of School of Agricultural Sciences, National Open University of Nigeria, the researcher may choose to observe and record it himself (primary data) or if the school keeps the records, he can use such data (secondary data).

Data could also be classified as **discrete** and **continuous** data. Discrete data have distinct value with no intermediate points. Examples of discrete data are household size, population census, number of students in a class, number of lecturers in a department, etc. The value of a discrete data must be in a whole number without fractions. Continuous data can take any value. It could be a whole number or fraction. For example, continuous data are weight, age, height, yield of crops, etc.

SELF-ASSESSMENT EXERCISE

i. What are data?
ii. What are the various ways in which data are classified?

4.0 CONCLUSION

In this unit, you have learnt the meaning of data and various techniques of data collection. Some of the basic concepts in this unit will be discussed at greater length in subsequent units. You have also learnt methods of classifying data in this unit. For the purpose of data collection, various methods can be employed in data collection.

5.0 SUMMARY

The main points in this unit include the following:

- Raw data need to be processed by subjecting it to some statistical analysis before it can make meaningful interpretation.
- Data can be collected from respondents through the use of structured questionnaire, phone or email.
- Data can be classified into quantitative and qualitative, primary or secondary and discrete or continuous data.

6.0 TUTOR-MARKED ASSIGNMENT

1. Briefly explain the different methods of data collection.
2. How can mass of data be classified?

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). VIKAS Publishing House PVT Limited, New Delhi.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 2 DATA PRESENTATION AND TABULATION

CONTENTS

1.0 INTRODUCTION

This unit is a continuation of module 2. In this unit, you will learn how to present data in different forms such as graphical presentation of data or tabular form. After studying this unit, you are expected to have achieved the objectives listed below.

2.0 OBJECTIVES

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

- outline briefly, the different methods of data presentation
- explain vividly how data is presented in a tabular form
- show practically how data is presented in a graphical form
- explain how data is presented in a mathematical form
- differentiate between grouped and ungrouped data.

3.0 MAIN CONTENT

3.1 Tabular Presentation of Data

In the presentation of data, the data need to be arranged in an array. An array is the arrangement of data (observations of a variable) either in an ascending or descending order of magnitude. Data is presented as either ungrouped or grouped data. Data organised and summarised in a frequency distribution are called grouped data while those that have not been organised in a frequency distribution are called ungrouped data.
Data can be presented either in a tabular form or in a chart. Grouped data are usually presented in a table where the frequency distribution is generated. Ungrouped data can be arranged in an array without putting it

in a table. Example, The following are the ages of National Open University of Nigeria Postgraduate Students:

24, 34, 31, 24, 24, 30, 17, 32, 28, 36, 40, 25, 44, 25, 36, 27, 40, 34, 28, 43,
This ungrouped data can be arranged in an array as:

17, 24, 24, 24, 25, 25, 27, 28, 28, 30, 31, 32, 34, 34, 36, 36, 40, 40, 43, 44.From here, we can determine the mean, the median and the mode directly from this variable. The formula for the mean = $\sum \frac{Xi}{n}$, where Xi is the individual score, n is the total number of observations and $\sum$ = summation sign. Therefore the mean = 622/20 = 31.1. The median is therefore the sum of the two middle values and divide by 2 because the number of observations is even. The median is therefore $\frac{30+31}{2}$ =305.

The mode of this distribution is the observation that repeated itself most (number with the highest frequency). The mode is therefore = 24.

Frequency distribution

When masses of raw data are summarised, it is important to distribute the data into categories called classes. In determining the number of individuals belonging to each class or frequency, a class width is usually developed. A tabular representation of data by classes including the corresponding frequencies is referred to as **frequency distribution.** Frequency distribution is made from an array so as to condense the data by reducing the number of rows such that closely related values are grouped onto class intervals. The following data above can be placed in a frequency distribution table as shown in Table 1.

Table 1: Ages of National Open University of Nigeria Postgraduate Students

Class interval	17-20	21-24	25-28	29-32	33-36	37-40	41-44
Frequency	1	3	5	3	4	2	2

Class interval and class limits

In Table 1, 17 - 20 in the first column is called class interval. The class intervals made up the class limits (lower and upper class limits). For example, in the class interval of 17-20, 17 is the lower class limit while 20 is the upper class limit. The class limits are the smallest (lower class limit) and the largest (upper class limit) observation in each class.

Class boundaries and class mark of a class interval

Class boundaries are obtained by adding the upper limit of one class interval to the lower limit of the next higher class interval and dividing by 2. Just as we have the upper and lower class limits, we also have lower and upper class boundaries. For example, in the second column of the Table 1 above, the lower class limit is 21 and the upper class limit is 24. The lower class boundary is the upper class limit of the class just before it plus the lower class limit of that class and dividing by 2. That is $\frac{20+21}{2}$ = 205. The upper class boundary of column 2 of Table 1 is the upper limit of that class plus the lower limit of the next class limit. That is $\frac{24+25}{2}$ = 24.5.

Class Mark: Class mark is the mid- point of the class interval obtained by adding up the lower and upper limits of each class and dividing by 2. For example, the class mark for column 1 in Table 1 = $\frac{17+20}{2}$ = 18.5. Class mark for column 2 = $\frac{21+24}{2}$ = 22.5 and so on.

Class Width: The width or size of a class is the difference between one lower class limit and the next lower class limit or difference between one upper class limit and the next upper class limit. From Table 1, the width of the class = 21 – 17 or 24 – 20, which is 4. The width of the class can also be determined by subtracting the lower limit from the upper limit of one class interval and add 1 to it. For example, the class interval of each class in Table 1 is 3 + 1 = 4.

Cumulative frequency and relative frequency

Cumulative frequency is the number of observations in that class plus all the observations above it. It indicates the total frequency of a set of class intervals. Relative frequency is also known as percentage. It is used for comparing the frequencies in each class relative to the total frequency and multiply by 100. It is the frequency of the class divided by the total frequency of all classes and is expressed as a percentage. From Table 1, the class boundary, class mark, frequency and relative frequency can be generated as shown in Table 2.

Table 2: Cumulative Frequency and Relative Frequency Table

Class interval	Class boundary		Class mark (X)	Frequency	Cumulative frequency	Relative frequency
17-20	16.5 20.5	–	18.5	1	1	5
21-24	20.5 24.5	–	22.5	3	4	15
25-28	24.5 28.5	–	26.5	5	9	25
29-32	28.5 32.5	–	30.5	3	12	15
33-36	32.5 36.5	–	34.5	4	16	20
37-40	36.5 40.5	–	38.5	2	18	10
41-44	40.5 44.5	–	42.5	2	20	10
$\sum$				**20**		**100**

3.2 Graphical Presentation of Data Histogram

A histogram consists of rectangles and vertical bars or columns whose heights are proportional to the frequencies in each class interval. The horizontal axis (X axis) represents the class boundaries to ensure that the bars are not separated. The vertical axis (Y axis) represents the frequency of the observations

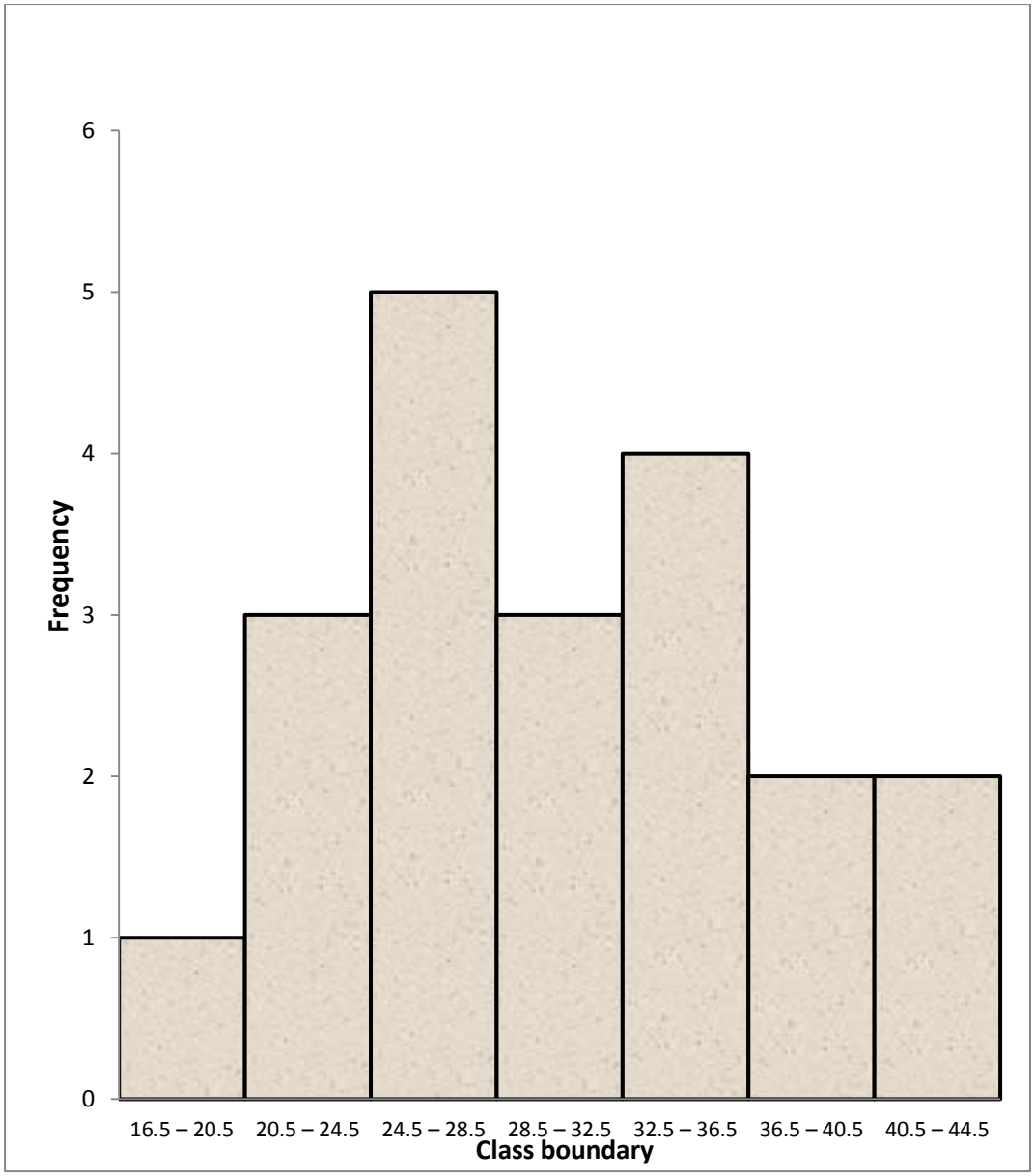


Figure 1: A Histogram showing the ages of NOUN Postgraduate students

A frequency polygon

A frequency polygon is a line graph joining the mid-point of the class interval. In a frequency polygon, a class mark is located in each class interval or class boundary. The horizontal axis (X axis) of the chart is made up of the class mark (instead of class boundary as obtained in histogram) while the vertical axis (Y axis) is made up of frequencies. A line is normally made to join all the points together. This curve is called frequency curve or frequency polygon.

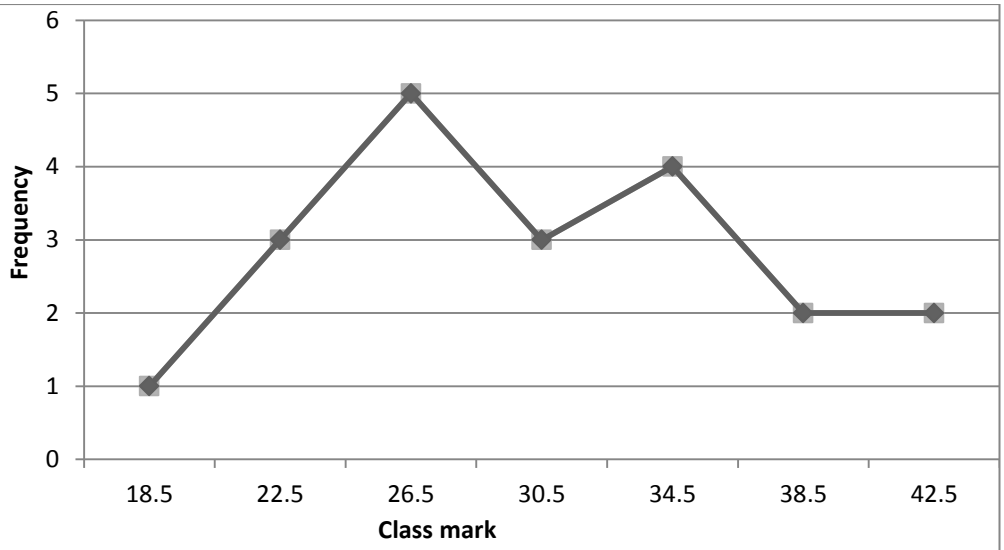


Figure 2: A frequency curve showing ages of Ahmadu Bello University Postgraduate students

Cumulative Frequency Polygon (Ogive)

A cumulative frequency curve or polygon can be formed by plotting the cumulative frequencies of the distribution on the vertical axis (Y axis) against the upper boundary on the horizontal axis (X axis). From Table 1, cumulative frequency polygon is plotted by using cumulative frequency values on the vertical axis and upper class boundary values on the horizontal axis as in Figure 3.

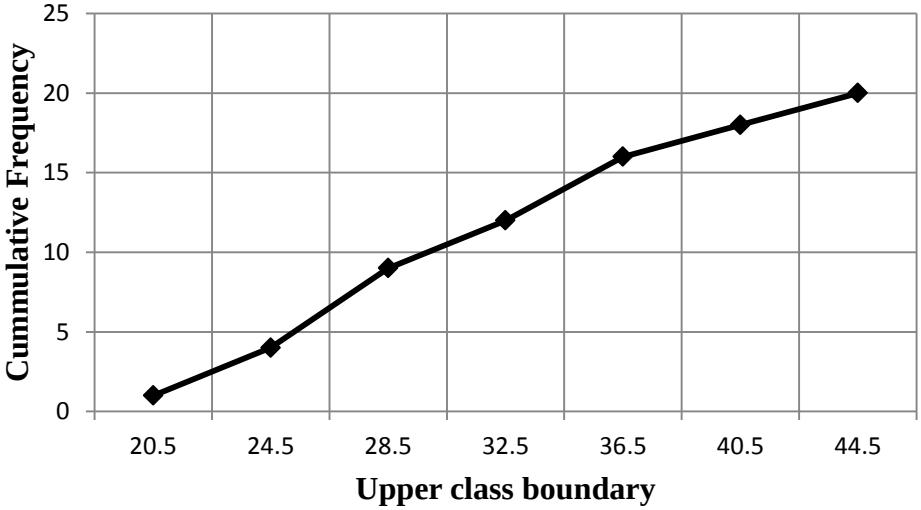


Figure 3: A cummulative frequency curve showing ages of NOUN Postgraduate students

Bar and Pie Charts

A bar chart is a graphical display of frequency distribution in bars. The bars can be vertical or horizontal, which are divided into sections. Each section of a bar chart corresponds in size to the frequency of the item it represents. A space is usually left between each adjacent bar so that the categories are separated. For example, the following are the values of Nigerian export of ground nut (₦ Million) between 2001 and 2005.

Table 3: Value of Nigerian Export of Groundnut between 2001 and 2005 (₦ Million)

Value of Nigerian export of groundnut (₦ Million) between 2001 and 2005	67	52	61	64	80	324
Year	2001	2002	2003	2004	2005	Total

This table can be plotted on a bar chart as shown in Figure 4.

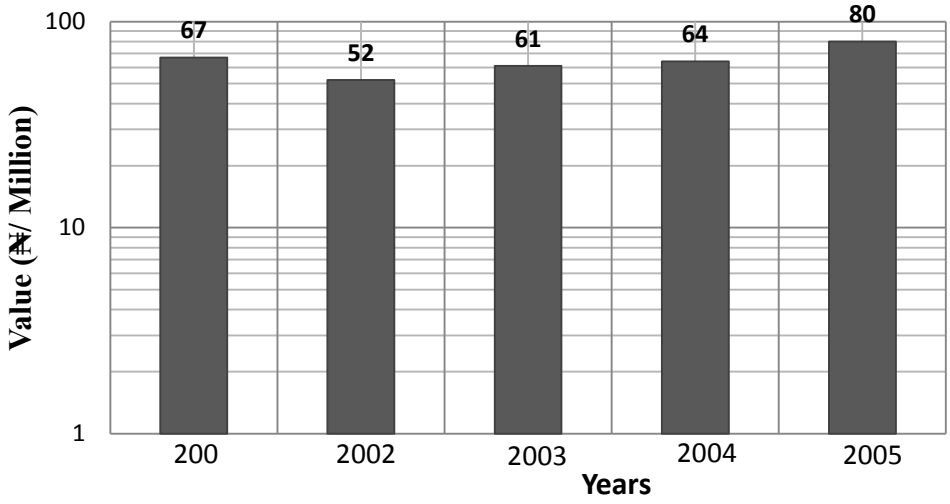


Figure 4: A bar chart showing the value of Nigerian export of groundnut (₦/ Million)

A pie chart

A pie chart is a circular diagram with the circumference of 360^0, in which each item is represented by a sector whose area is proportional to the percentage or relative frequency of the total. In the drawing of a pie chart, the following steps are followed:

1. Find the angle of the sector corresponding to each item in a circumference of 360^0. The formula for finding this is
$$\frac{Fi}{\sum Fi} \text{X } 360$$
Where: Frequency or value contributed by each item and $\sum Fi$ = the sum of all the frequencies or total value of the items.

2. Find the relative frequency or percentage value of each item in relation to the total value of items. The formula for finding this is
$$\frac{Fi}{\sum Fi} \text{X } 100$$

Using Table 3 above, the value of Nigerian export of groundnut can be represented on a pie chart as shown in Figure 5. This can be calculated as shown in Table 4.

Table 4: Value of Nigerian Export of Groundnut (2001 -2005) and their Degree and Percentage Contribution

Year	2001	2002	2003	2004	2005
Value (₦/ Million)	67	52	61	64	80
Degree contributed	$\frac{67}{324} \text{ x } 360 = 74.44^0$	$\frac{52}{324} \text{ x} 360 = 57.78^0$	$\frac{61}{324} \text{ x } 360 = 67.78^0$	$\frac{64}{324} \text{ x } 360 = 71.11^0$	$\frac{80}{324} \text{ x } 360 = 88.89^0$
Percentage contribution	$\frac{67}{324} \text{ x } 100 = 20.68$	$\frac{52}{324} x 100 = 16.05$	$\frac{61}{324} \text{ x } 100 = 18.83$	$\frac{64}{324} \text{ x } 100 = 19.75$	$\frac{80}{324} \text{ x } 360 = 24.69$

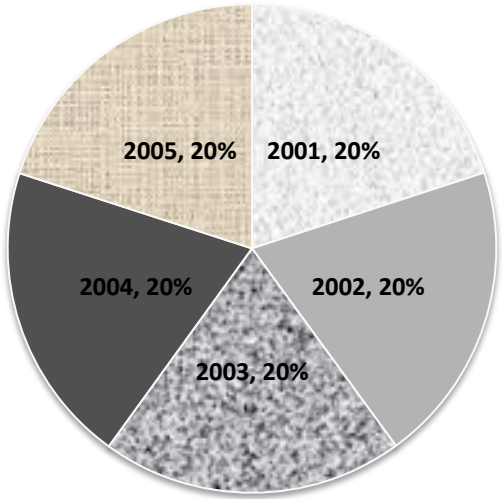


Figure 5: A pie chart showing the value of Nigerian export of groundnut (2001-2005)

SELF-ASSESSMENT EXERCISE

i. What is an array?
ii. Use the following data below to solve the following problems:

24, 34, 31, 24, 24, 30, 17, 32, 28, 36, 40, 25, 44, 25, 36, 27, 40, 34, 28, 43,

a. Make a frequency distribution of this data and find the class interval, class boundary, frequency and cumulative frequency.
b. Draw the frequency curve and cumulative frequency polygon.

4.0 CONCLUSION

This unit has introduced you to different methods of presenting data as well as various forms of graphical presentations. This unit centred mainly on data presentation in various forms. The graphical presentation of data includes histogram, line graph, bar charts and pie charts.

5.0 SUMMARY

The main points in this unit include the following:

- Ungrouped data can be grouped using frequency distribution.
- Statistical data can be presented in tabular form or graphically.
- Data can be presented graphically as a histogram, line graph, bar chart or pie chart.

6.0 TUTOR-MARKED ASSIGNMENT

Briefly explain the following terms:

1. class mark
2. class width
3. histogram
4. frequency polygon
5. cumulative frequency polygon
6. a bar chart
7. a pie chart.

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Gupta, C. B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

MODULE 3 MEASURES OF CENTRAL TENDENCY

Unit 1 Measures of Central Tendency (The Arithmetic, Weighted, Geometric and Harmonic Mean)
Unit 2 Measures of Central Tendency (Median)
Unit 3 Quartile and Percentile
Unit 4 The Mode and Relationship Between Mean, Median and Mode

UNIT 1 MEASURES OF CENTRAL TENDENCY (THE MEAN)

CONTENTS

1.0 INTRODUCTION

Measures of central tendency especially the mean is highly useful in our daily research work. The different means that will be treated in this unit are also measures of central tendency in the sense that the value tends to lie centrally within a set of data arranged according to magnitude. The arithmetic mean, weighted arithmetic mean, geometric mean and the harmonic mean are the topic of discussion for this unit.

2.0 OBJECTIVES

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

- define the different types of statistical means
- differentiate between arithmetic, weighted, geometric and harmonic means.

3.0 MAIN CONTENT

3.1 Measures of Central Tendency (The Mean)

Measures of central tendency can be described as a number or value that is representative of other numbers or values by bringing the set of data towards the centre. The popular measures of central tendency are the mean or arithmetic mean, median and mode. Averages are also measures of central tendency since the value tend to lie centrally within a set of data arranged according to magnitude.

1. Arithmetic mean

The arithmetic mean or simply called the mean is the sum of all the observations divided by the number of observations in the sample. The arithmetic mean is usually denoted by 'X bar' ($\overline{X}$). For example, if X is the variable that takes values X1, X2, X3....Xn on n items, then the arithmetic mean is defined as

$$\overline{X} = \frac{X_1 + X_2 + X_3 ..... X_n}{n} = \frac{\sum X_i}{n}$$

Where:

$\sum$ = summation sign

X_1 - X_n =Value of items

X_i = Value of individual Xs

N = Number of items (sample size).

Example 1. The arithmetic mean of numbers 10, 6, 18, 20 and 22 is:

$$\overline{X} = \frac{\sum X_i}{n} = \frac{10+6+18+20+22}{5} = \frac{76}{5} = 15.2$$

The formula above is employed when the data are not grouped, if the data are grouped (with frequencies), the mean is

$$\overline{X} = f_1 X_1 + f_2 X_2 + f_3 X_3 + ... f_n X_n = \frac{\sum fX}{\sum f}$$

Where:

$\sum fX$ =Summation the frequencies multiplied by X in each case

$\sum f$ = Total frequency

Example 2: Table 5 below gives the frequency distribution of the average monthly earnings of male workers. Calculate the mean earnings (Gupta and Gupta, 2010).

Table 5: Frequency Distribution of the Average Monthly Earnings of Male Workers

Monthly earnings (Rs)	Midpoint (m)	No. of workers (f)	f m
27.5 – 32.5	30	120	3,600
32.5 – 37.5	35	152	5,320
37.5 – 42.5	40	170	6,800
42.5 – 47.5	45	214	9,630
47.5 – 52.5	50	410	20,500
52.5 – 57.5	55	429	23,590
57.5 – 62.5	60	568	34,080
62.5 – 67.5	65	650	42,250
67.5 – 72.5	70	795	55,650
72.5 – 77.5	75	915	68,625
77.5 – 82.5	80	745	59,600
82.5 – 87.5	85	530	45,050
87.5 – 92.5	90	259	23,310
92.5 – 97.5	95	152	14,440
97.5 – 102.5	100	107	10,700
102.5 – 107.5	105	50	5,250
107.5 – 112.5	110	25	2,750
$\sum$		**6,291**	**431,150**

Solution

Since the variable is a continuous one, the midpoints are calculated simply as (lower limit + upper limit) /2

The mean is calculated as

$$\overline{X} = \frac{\sum fm}{\sum f} = \frac{431150}{6291} = \text{Rs } 68.5$$

2. Weighted arithmetic mean

This is a situation where we associate with the values X_1, X_2, X_3...X_n certain weights which could be w_1, w_2, w_3....w_n, depending on the importance attached to the weights. It is a set of mean weighted by the number of cases it is based on. The weighted arithmetic mean can be calculated using the formula:

$$\overline{X} = \frac{w_1X_1 + w_2X_2 + w_3X_3 + ...w_nX_n}{w_1 + w_2 + w_3....w_n} = \frac{\sum wX}{\sum w}$$

Where:

$\overline{X}$ =weighted arithmetic mean

X_1 - X_n = Values of X_1 - X_n respectively

w_1 w_n = weights attach to X_1 - X_n respectively

Example: In 2012, three terminal examinations were conducted for NOUN staff school children. If the second term score of one of the candidate was weighted twice as much as first and third

term exam was thrice as much as his first. Calculate the weighted arithmetic mean of the child if the candidate has 80, 75 and 70 in his first, second and third term respectively.

Solution

$$\overline{X} = \frac{1(80)+2(75)+3(70)}{1+2+3} = \frac{80+150+210}{6} = \frac{440}{6} = 73.33$$

Example 2: The arithmetic mean of daily wages of two manufacturing concerns A Ltd. And B Ltd is Rs 5 and Rs 7 respectively. Determine the average daily wages of both concerns if the number of workers employed were 2,000 and 4,000 respectively (Gupta and Gupta, 2010)

Solution

The following procedures are followed in answering this question:

i. Multiply each mean (5 and 7) by the number of workers in the concern it represents
ii. Add up the two products obtained in (i) above
iii. Divide the total obtained in (ii) by the total number of workers.

The arrangement is presented in Table 6 as below:

Table 6: Weighted Mean of Mean Wages of A Ltd and B Ltd.

Manufacturing concern	Mean wages (X)	Workers employed (w)	Mean wages x workers employed
A Ltd	5	2,000	10,000
B Ltd	7	4,000	28,000
Total	12	6,000	38,000

$$\overline{X} = \frac{\sum wX}{\sum w} = \frac{38,000}{6,000} = \text{Rs } 6.33.$$

3. Geometric mean

The geometric mean of n positive values is defined as the nth root of the product of the values. Instead of adding the observations and dividing by the total number, the observations are multiplied and the nth root of the product is the geometric mean. The geometric mean of a set of n positive numbers $X_1, X_2, X_3 ..... X_n$ is calculated as:

GM = $\sqrt[n]{X_1, X_2, X_3 ..... X_n}$. The geometric mean is always less than the arithmetic mean.

Where:
n = Total number of items
Example 1: The geometric mean of 2, 4, 8 and 64 is:
$\sqrt[4]{2 \text{ x } 4 \text{ x } 8 \text{ x } 64} = \sqrt[4]{4096} = 8$
Geometric mean is also used in calculating mean of ratios.
Example 2: The geometric mean of the fractions $\frac{2}{3}, \frac{2}{5}$ and $\frac{3}{5}$ is

$$\frac{\sqrt[3]{2 \text{ x } 2 \text{ x } 3}}{\sqrt[3]{3 \text{ x } 5 \text{ x } 5}} = \frac{\sqrt[3]{12}}{\sqrt[3]{75}} = \frac{2.29}{4.22} = 0.54$$

This method is easily applicable if when then sample size (n) is small, i.e. $n \leq 4$. Where n is large, the method cannot be comfortably applied. In order to simplify the computation of geometric mean, logarithms are employed. For ungrouped data,

Logf GM = $\frac{\log X_1 + \log X_2 + l\log X_3 ..... \log X_n}{n}$

Example 1: The geometric mean of 3, 4, 5, 8, 11, 12, 14, and 16 is:

Table 7: Table of Values and their Log

Values	**log**
3	0.4771
4	0.6021
5	0.6990
8	0.9031
11	1.0414
12	1.0792
14	1.1461
16	1.2041
Total	**7.1521**

GM = $\frac{7.1521}{8}$ =0.894
Antilog of 0.894 = 7.834
GM = 7.834.
For grouped data, the geometric mean of a frequency distribution is given by

Log GM = $\frac{\sum(f \log X)}{\sum f} = \frac{f_1 \log X_1 + f_2 \log X_2 + f_3 \log X_3 .... f_n \log X_n}{\sum f}$

Where:
X = Mid value of a particular class
F = Frequency of each class

$\sum f$ = Total frequency

Example 2: The following is the household expenditure for a particular month in thousand naira per day.

Table 8: Household Expenditure in Thousand Naira Per Day

Expenditure	No. of families (f)	Mid value (X)	Log X	F x (log X)
0.5 -11.5	8	6	0.7782	6.2256
11.5 – 20.5	12	16	1.2041	14.4492
21.5 – 30.5	14	26	1.4150	19.8100
31.5 – 40.5	10	36	1.5563	15.5630
41.5 – 50.5	6	46	1.6628	9.9768
Total	**50**			**66.0246**

$$GM = \frac{\sum(f \log X)}{\sum f} = \frac{66.0246}{50} = 1.3205$$

Antilog of 1.3205 = 20.91
GM = 20.91

4. Harmonic mean

The harmonic mean (HM) of a set of n numbers is the reciprocal of the arithmetic mean of the reciprocal of the numbers.

$$HM = \frac{n}{\sum\left(\frac{1}{Xi}\right)} = \frac{n}{\left(\frac{1}{X_i} + \frac{1}{X_2} + \frac{1}{X_3} \ldots \frac{1}{X_n}\right)}$$

Where:
n = sample size (number of items or observation)
X_i= Value of items X_1, X_2, X_3...X_n
HM = Harmonic Mean

Therefore, $\frac{1}{HM} = \frac{\sum\left(\frac{1}{X_i}\right)}{n}$

$$\frac{1}{HM} = \frac{\left(\frac{1}{X_i} + \frac{1}{X_2} + \frac{1}{X_3} \ldots \frac{1}{X_n}\right)}{n}$$

Example 1: The harmonic mean of the numbers 4, 8 and 10 is:

$$HM = \frac{3}{\frac{1}{4} + \frac{1}{8} + \frac{1}{10}} = \frac{3}{19/40}$$

$$= 6.316$$

The second formula can also be applied as:

$$\frac{1}{HM} = \frac{\frac{1}{4}+\frac{1}{8}+\frac{1}{10}}{3} = \frac{19/40}{3} = \frac{0.475}{3}$$

$\frac{1}{HM} = \frac{0.475}{3}$. Then HM = $\frac{3}{0.475}$ = 6.316

Example 2: If a motorcyclist travels 150km each day for five days at speeds of 20, 30, 40, 50 and 60km/h respectively, his average speed can be calculated using the harmonic mean as:

$$HM = \frac{5}{\frac{1}{20}+\frac{1}{30}+\frac{1}{40}+\frac{1}{50}+\frac{1}{60}} = \frac{5}{87/600} = \frac{5}{0.145}$$

= 34.48km/h

The average speed for the five days using the harmonic mean is 34.48km/h. The harmonic mean gives the correct average speed than just the arithmetic mean because the speed differs in each day.

Note: The harmonic mean is always less than the geometric mean and the arithmetic mean. The sequence of their magnitude is:
AM > GM > HM
Where:

AM = Arithmetic mean
GM = Geometric mean
HM = Harmonic mean

SELF-ASSESSMENT EXERCISE

If the wage paid to five staff of a company on a certain year is 3, 4, 5, 8 and 10 Find:

i. The arithmetic mean
ii. The geometric mean
iii. The harmonic mean of their wages

4.0 CONCLUSION

This unit exposed us to the calculation of different statistical means (arithmetic mean, weighted mean, geometric mean and harmonic mean). These statistical means are part of measures of central tendency because they are representative of other numbers or values by bringing the set of data towards the centre.

5.0 SUMMARY

The main points in this unit are as follows:

- The arithmetic mean or the mean is the sum of all the observations divided by the number of observations in the sample.
- Arithmetic mean can be used for both grouped and ungrouped data
- Weighted arithmetic mean is a situation where we associate with the values X_1, X_2, X_3...X_n certain weights which could be w_1, w_2, w_3....w_n, depending on the importance attached to the weights
- Geometric mean of n positive values is defined as the nth root of the product of the values.
- Where the sample size n is large, the computation of geometric mean is simplified by employing logarithms.
- Geometric mean can be employed for both grouped and ungrouped data.
- The harmonic mean (HM) of a set of n numbers is the reciprocal of the arithmetic mean of the reciprocal of the numbers.

6.0 TUTOR-MARKED ASSIGNMENT

1. The mark of 15 NOUN students given in Table 9 is as shown below: Calculate their mean mark.

Table 9: Mark of 15 NOUN Students

Marks	Frequency
10	3
12	2
16	4
18	1
20	5
Total	**15**

2. If a final examination in a course is weighted three times as much as a quiz and a student has a final examination grade of 85 and quiz of 70 and 90, find the mean grade (Spiegel and Stephens (1999).

7.0 REFERENCES/FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Brink, D. (2010). *Essentials of Statistics.* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics*. Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

UNIT 2 MEASURES OF CENTRAL TENDENCY (MEDIAN)

CONTENTS

1.0 INTRODUCTION

From the discussion of last unit (unit 1), you can calculate the arithmetic mean, weighted, geometric and harmonic mean. The topic of discussion in this unit is the calculation of median from both grouped and ungrouped data. After studying this unit, you are expected to have achieved the stated objectives of this unit.

2.0 OBJECTIVES

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

- determine the median from both grouped and ungrouped data
- use cumulative frequency polygon (ogive) to determine the median of the distribution.

3.0 MAIN CONTENT

3.1 The Median

The median of a set of observations arranged in an array is either the middle value or the mean of the two middle values when the number of observations is even. It is the second measure of central tendency that has wide application in statistical works. For grouped data, the median can easily be determined if the observations are arranged either in ascending or descending order of magnitude. For ungrouped data, the median can be determined as shown in the example below.

Example: The set of numbers; 24, 34, 31, 24, 24, 30, 17, 32, 28, 36, 40, 25, 44, 25, 36, 27, 40, 34, 28, 43 and 42.

The set of 21 observations can be arranged in an array as:
17, 24, 24, 24, 25, 25, 27, 28, 28, 30, 31, 32, 34, 34, 36, 36, 40, 40, 42, 43, 44.

There are 21set of observations here, the middle value is definitely the 11^{th} observation, counting either from the left or right. The 11^{th} observation is 31, therefore, the median of this distribution is 31. The middle value is chosen because the number of observations is odd. Assuming the last value (42) before arranging in an array is removed, the remaining observation will be 20 which is an even number as shown below:

17, 24, 24, 24, 25, 25, 27, 28, 28, 30, 31, 32, 34, 34, 36, 36, 40, 40, 43, 44.
Here, the median is the mean of the 10^{th} and the 11^{th} numbers, which is 30 and 31.

Thus, the median = $\frac{30+31}{2} = 30.5$

For grouped data, the median of the classes can also be determined. There are two methods of determining the median score from grouped data:

Method 1

Prepare the cumulative frequency of the distribution

1. Divide the total frequency by two and locate the class with this value.
2. Draw the ogive.
3. Draw horizontal line at the point of the value obtained in step 2 until it meets the curve.
4. Draw a vertical line from the curve to the X axis and read the value of X at this point.

The value of X is the median of the distribution.

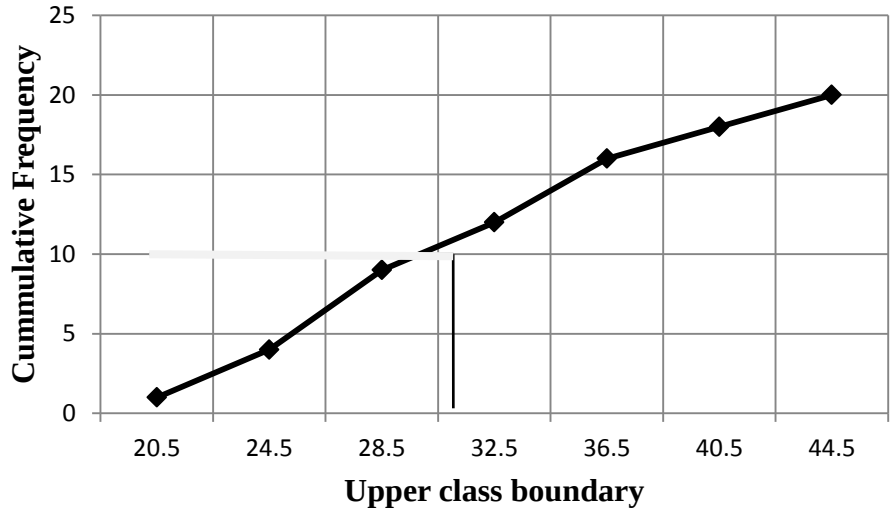


Figure 3: A cummulative frequency curve showing ages of NOUN Postgraduate students

From this curve, the value is 29.83.

Method 2

1. Prepare the cumulative frequency of the distribution.
2. Divide the total frequency by two and locate the value in the table.
3. Use the formula below to obtain the median class.

$$\text{Median} = L_1 + \left(\frac{\frac{\sum f}{2} - \sum f_0}{f_1} \right) \text{x C}$$

Where:

L_1= Lower class boundary of the median class

$\sum f$ = Total frequency

$\sum f_0$ = Sum of the frequencies of all classes just before the median class

f_1 = Frequency of the median class

C = Class width

Example: The above data for 20 observations can be made in a frequency distribution table as shown in Table 10.

Table 10: A Frequency Distribution Table

Class interval	Class boundary	Class mark (X)	Frequency (f)	Cumulative frequency (Cf)
17 – 20	16.5 – 20.5	18.5	1	1
21 – 24	20.5 – 24.5	22.5	3	4
25 – 28	24.5 – 28.8	26.5	5	9
29 – 32	28.5 – 32.5	30.5	3	12
33 – 36	32.5 – 36.5	34.5	4	16
37 – 40	36.5 – 40.5	38.5	2	18
41 – 44	40.5 – 44.5	42.5	2	20
Total			**20**	

$$\text{Median} = L_1 + \left(\frac{\frac{\sum f}{2} - \sum f_0}{f_1} \right) \text{x C}$$

$$= 28.5 + \left(\frac{10 - 9}{3} \right) 4 = 28.5 + \left(\frac{1}{3} \right) 4 = 29.83$$

SELF-ASSESSMENT EXERCISE

The number of ATM transactions per day was recorded at 15 locations in a large city. The data were: 35, 49, 225, 50, 30, 65, 40, 55, 52, 76, 48, 325, 47, 32 and 60. Find the median of this observation (Spiegel and Stephens, 1999).

4.0 CONCLUSION

In this unit, you have learnt to calculate median from both grouped and ungrouped data. You have also learnt how to use cumulative frequency curve (ogive) to determine the median of any data.

5.0 SUMMARY

The main points in this unit include the following:

- The median of a set of observations arranged in an array is either the middle value or the mean of the two middle values when the number of observations is even.
- Median of both grouped and ungrouped data can be determined.
- Cumulative frequency curve or polygon can be used to determine the median of a frequency distribution.

6.0 TUTOR-MARKED ASSIGNMENT

1. Given the following age distribution of NOUN postgraduate students as shown in Table 10 b. Find the median of this distribution.

Table 10b: Age Distribution of NOUN Postgraduate Students

Classes	Frequency
29 – 35	5
36 – 42	2
43 - 49	3
50 – 50	1
57 – 63	3
64 - 70	1
Total	**15**

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 3 QUARTILES AND PERCENTILES

CONTENTS

1.0 INTRODUCTION

Generally, the median formula is applied in both quartile and percentile, depending on the definition of N. First quartile = $^{1N}\!/_{4}$, second quartile = $^{2N}\!/_{4}$ = $^{N}\!/_{2}$ = Median Third quartile = $^{3N}\!/_{4}$. First percentile = $\frac{1N}{100}$, 50^{th} percentile = $\frac{50N}{100} = \frac{N}{2} = \frac{\sum f}{2}$ = Median and 90^{th} percentile = $\frac{90N}{100}$. At the end of this unit, you are expected to have achieved the stated objectives.

2.0 OBJECTIVES

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

- define and explain how quartile is calculated
- differentiate between quartile and percentile.

3.0 MAIN CONTENT

3.1 Quartile

The nth quartile is the value of the item which lies at the n $\left(^{N}\!/_{4}\right)$th item. Generally, the median formula is applied in both quartile and percentile, depending on the definition of N.

First quartile = $^{1N}/_{4}$, second quartile = $^{2N}/_{4} = {}^{N}/_{2} =$ Median

Third quartile = $^{3N}/_{4}$.

Example: Find the first, second and third quartile of the distribution in Table 10

Solution:

First quartile $Q_1$is the value of $\frac{N}{4}$ of the cumulative frequency.

N = 20, then $\frac{N}{4} = 5$. The appropriate class boundary is 24.5 – 28.8.Now we can determine the lower class boundary of the Q_1using the median formula.

$$Q_1 = L_1 + \left(\frac{\frac{\sum f}{4} - \sum f_0}{f_1} \right) x\, C$$

Where:

Q_1= First quartile of the distribution

L_1= Lower class boundary of the first quartile class

$\sum f$ = Total frequency

$\sum f_0$ =Cumulative frequency of the class just before the Q_1 class

f_1 = Frequency of the Q_1 class

C = Class width

$$Q_1 = 24.5 + \left(\frac{5-4}{5} \right) 4$$

= 24.5 + 0.8 = 25.30

Second quartile $Q_2 = \frac{2N}{2} = \frac{\sum f}{2} = \text{Median}, N = 10$

From our former result, the median = 29.83

Third quartile $Q_3 = \frac{3N}{4} = \frac{3x20}{4} = \frac{60}{4} = 15$. The appropriate class boundary is 32.5 – 36.5. We can now determine the third quartile as:

$$Q_3 = 32.5 + \left(\frac{15-12}{4} \right) 4$$

Q_3= 32.5 + 3 = 35.5

3.2 Percentile

The nth percentile of observations arranged in an array is the value of the item lie below it.

n Percentile (P_n) = $L1 + \left(\frac{N/100 - Cf_0}{f_1} \right)$ x C

Where:

P_n= n percentile

L1 = Lower class boundary of the n percentile class

N = $\sum f$ = Total frequency

Cf_0= Cumulative frequency of the class just before the n percentile class

f_1 = Frequency of the n percentile class

C = Class width

The first percentile = $\frac{1N}{100} = \frac{\sum f}{100}$

50^{th} percentile = $\frac{50N}{100} = \frac{N}{2} = \frac{\sum f}{2}$ = Median

90^{th} percentile = $\frac{90N}{100}$

Example: Find the 10^{th} and 90^{th} percentile of the data given in Table 10.

Solution:

From Table 10, 10^{th} percentile = $\frac{10N}{100} = \frac{\sum f}{10}$

$\sum f = 20$, so $\frac{\sum f}{10} = 2$.

Therefore, the class boundary that falls in this category is 20.5 – 24.5.

10^{th} percentile (P_{10}) = $20.5 + \left(\frac{2-1}{3} \right) 4$

= 20.5 + 1.33 = 21.83

90^{th} percentile = $P_{90} = \frac{90N}{100} = 18$. The lower class boundary that falls in this category is 36.5.

P_{90}= $36.5 = \left(\frac{18-16}{2} \right) 4$ =36.5 + 4

= 40.5.

SELF-ASSESSMENT EXERCISE

From Table 10, calculate the 50th percentile.

4.0 CONCLUSION

In this unit you have learnt how to calculate quartile and percentile from any frequency distribution. The calculation of first, second and third quartile have been demonstrated as well as 10th, 50th and 90th percentile. From this discussion, you would now be able to calculate the various types of quartiles and percentiles.

5.0 SUMMARY

Here are the major points in this unit:

- The median formula is applied in both quartile and percentile, depending on the definition of N.
- 2nd quartile and 50th percentile are equal to the median of the distribution
- The nth percentile of observations arranged in an array is the value of the item lie below it.

6.0 TUTOR-MARKED ASSIGNMENT

From Table 10b, find the second quartile, 10th percentile and 50th percentile.

7.0 REFERENCES/FURTHER EADING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 4 MODE

CONTENTS

1.0 INTRODUCTION

This unit will consider the mode as a measure of central tendency and the relationship between the different measures of central tendency (mean, median and mode). Understanding the whole of this module is very important because it is highly applicable in any statistical analysis particularly descriptive statistics.

2.0 OBJECTIVES

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

- define what mode is all about
- calculate mode from both grouped and ungrouped data
- use frequency curve or polygon to determine mode of a frequency distribution
- explain vividly how a normal distribution curve is skewed to the left of right.

3.0 MAIN CONTENT

3.1 The Mode

The mode of a set of data is the observation which occurs most frequently. In some set of data, the mode may not exist, and if it does, there may be two or more modes. If we have 1, 2, 3 or more modes in a distribution, it is called unimodal, bimodal, trimodal and multimodal respectively. Using the above data that consists of 20 observations:
17, 24, 24, 24, 25, 25, 27, 28, 28, 30, 31, 32, 34, 34, 36, 36, 40, 40, 43, 44.

The mode for this distribution is 24 because it is only 24 that appears three times.

For grouped data, the modal class is the class with the highest frequency. The mode can be estimated using two methods:

1. Draw a frequency polygon and locate the value at the highest point
2. Use the formula for mode as given below:

$$\text{Mode} = L1 + \frac{(F_1 - F_0)}{(2F_1 - F_0 - F_2)} \times C$$

Where:

L1 = Lower class boundary of the modal class

F_0 = Frequency of the interval just before the modal class

F_1 = Frequency of the modal class

F_2 = Frequency of the class just after the modal class

C = Class width

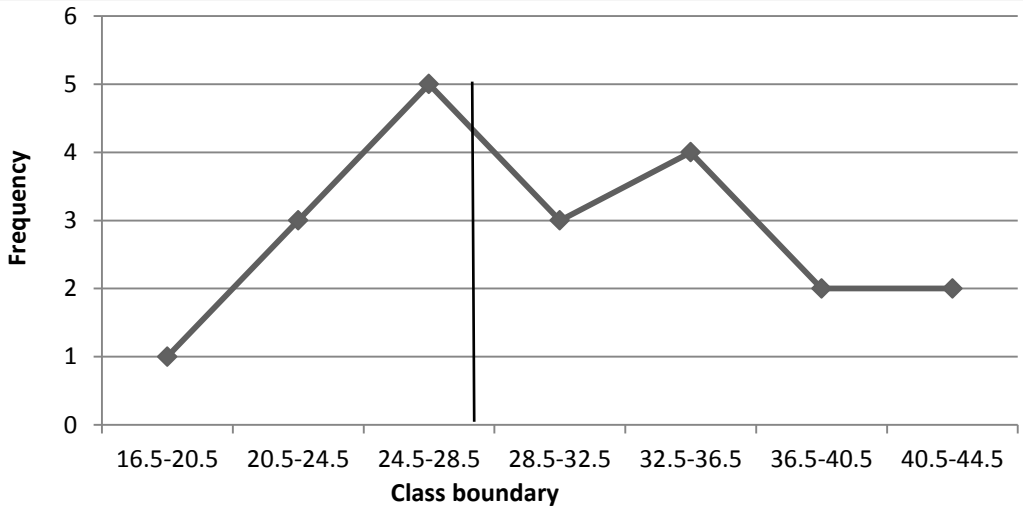
From Table 10c as shown below, the modal class can be obtained from the table using the above formula.

Class interval	**Class boundary**	**Class mark (X)**	**Frequency (f)**	**Cumulative frequency (Cf)**
17 – 20	16.5 – 20.5	18.5	1	1
21 – 24	20.5 – 24.5	22.5	3	4
25 – 28	24.5 – 28.8	26.5	5	9
29 – 32	28.5 – 32.5	30.5	3	12
33 – 36	32.5 – 36.5	34.5	4	16
37 – 40	36.5 – 40.5	38.5	2	18
41 – 44	40.5 – 44.5	42.5	2	20
Total			**20**	

$$\text{Mode} = 24.5 + \left(\frac{5-3}{10-3-3}\right)4 = 24.5 + \left(\frac{2}{4}\right)4$$

$$= 24.5 + 2 = 26.5.$$

From this data on Table 10, using the first method by drawing the frequency polygon (see figure 2) as in figure 2, the value of the highest point can be traced on the X axis. The highest point there is between 24.5 and 28.5. The average of these two values gives you 26.5. This value should exactly be 26.5 if it is correctly done.



From Figure 2: Determination of mode using a frequency curve

3.3 Relationship Between Mean, Median and Mode

When we have a distribution that has uniform number of high and low scores, the distribution is said to be normal. This means that the population mean, population median and the mode are all located at the centre. This can be represented by a symmetrical bell-shaped curve in which the curve can be divided into two equal halves (one half is a mirror image of the other).

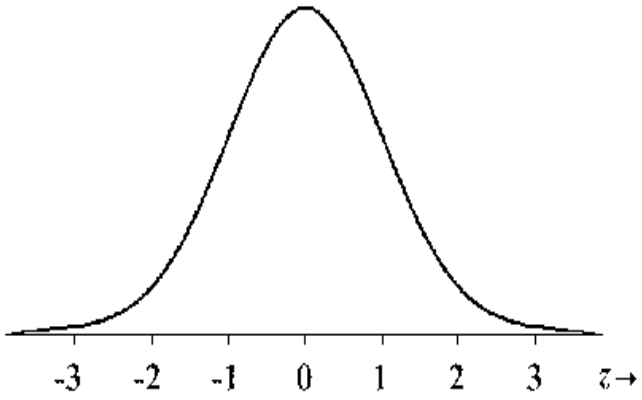
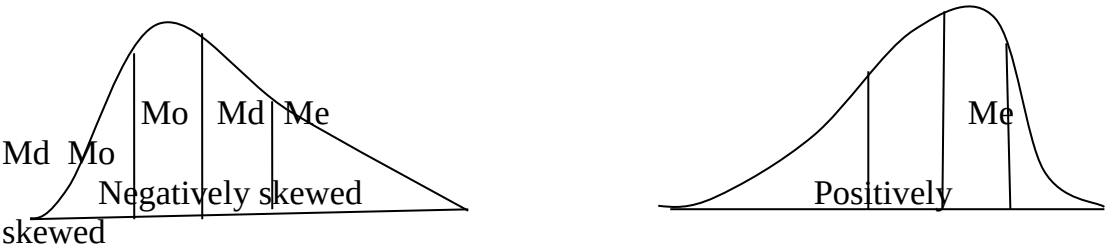


Figure 6: Standard Normal Distribution Curve

When the distribution has either larger number of relatively low or high scores, then the mean, median and mode have different values and the shape of the curve is asymmetrical. If the distribution is such a way that the curve is negatively skewed to the left, it means that the mean is the lowest while the mode is the highest. This implies that the set of

observation has larger number of relatively large scores where mode is relatively high. In this case,
Mode >Median > Mean



Note:
Mo = Mode
Md = Median
Me = mean

However, if the distribution has larger number of relatively low scores, then the mean has the largest value while the mode has the least. Such a situation is positively skewed to the right. This means that Mean > Median > Mode. In whichever the direction of skewness of the curve, the median always divide the curve into two.

SELF-ASSESSMENT EXERCISE

Using Table 10 b, find:

i. The mode of the distribution.
ii. How is this mode different from median and mean.

4.0 CONCLUSION

This unit has introduced you to the calculation of mode from both grouped and ungrouped data. Also, mode can be determined graphically, using a frequency polygon. Mean, median and mode are related in one way or the other. Where mode is greater than the mean, the curve is usually negatively skewed with longer tail to the right. Where mean is greater than the mode, the curve is usually positively skewed with longer tail to the left.

5.0 SUMMARY

The main points in this unit are:

- The mode is the number that occurs most frequently in a distribution
- Mode can be determined from both grouped and ungrouped data.

- Frequency polygon can be used to determine mode of a frequency distribution
- Mean, median and mode are related in one form or the other
- Due to the distribution of the scores in the data, the curve of this distribution may either be normal, negatively skewed or positively skewed.

6.0 TUTOR-MARKED ASSIGNMENT

1. Suppose the intelligent quotient (IQs) of 60 randomly selected NOUN students are presented in Table 10d as below:

Table 10 d: Intelligent Quotient (IQs) of 60 Randomly Selected NOUN students

IQs	69 - 72	73 - 76	77 - 80	81 - 84	85 - 88	89 - 92	93 - 96	97-100	101 - 104	105 - 108
Frequency	4	9	16	8	6	5	2	5	3	2

Using a class width of 4, find the mode of the distribution

7.0 REFERENCES/FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

MODULE 4 MEASURES OF DISPERSION

Unit 1 Measure of Dispersion (the Range and Mean Deviation)
Unit 2 Standard Deviation and Variance as a Measure of Dispersion

CONTENTS

UNIT 1 MEASURE OF DISPERSION (THE RANGE AND MEAN DEVIATION

1.0 INTRODUCTION

I believe you have understood the previous module which is centred on measures of central tendency. This module is centred on measures of dispersion and how it is applied in each case. The objectives below specify what you are expected to have learnt after studying this unit.

2.0 OBJECTIVES

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

- define what range is all about
- calculate the range from any given data
- determine the mean deviation from any frequency distribution.

3.0 MAIN CONTENT

3.1 Measure of Dispersion

The dispersion or variation is the degree to which numerical data tend to spread about an average value of the data. It is used to observe the degree of spread or unevenness of a set of observation. The most common measures of dispersion are range, mean deviation, standard deviation, variance and coefficient of variation.

The range: The range of a set of data is the difference between the largest and smallest number and observation in a set of data. Range only takes into consideration the extreme values in a set of data.

Example: The range of a set; 2, 3, 5, 6, 8, 12, 18 and 25 is: 25 – 2 = 23. The range is only affected by two extreme values in the entire distribution, any change or removal of any other observation would not affect it. For example, if there are 50 set of observations in a distribution and the lowest value is 5, the highest value is 100, even if about 46 observations are removed and it remains only 4 observations like 2, 4, 50 and 100. The range will still remain unchanged.

3.2 Mean Deviation

The mean deviation of a frequency distribution is the mean of the absolute values of the deviation from some measures of central tendency such as mean. Mean deviation is obtained by ignoring the signs of the deviation and simply regarding all of them as positive. The mean of these absolute deviations is referred to as mean deviation.

Steps in calculating mean deviation

1. Calculate the mean of the deviation ($\overline{X}$).
2. Record the deviation from the mean in each observation: d = $\left(X - \overline{X}\right)$
3. Convert the deviations to absolute value.
4. Find the average value of deviations (mean deviation).

For ungrouped data, mean deviation = $\frac{1}{n}\sum\left(X - \overline{X}\right) = \frac{\sum\left(X - \overline{X}\right)}{n}$

Example: From this ungrouped data: 17, 24, 24, 24, 25, 25, 27, 28, 28, 30, 31, 32, 34, 34, 36, 36, 40, 40, 43, 44. The mean is $\frac{\sum Xi}{n} = \frac{622}{20} = 31.1$

To find the mean deviation, the data can be arranged in a tabular form as shown in Table 11.

Table 11: A Mean Deviation Table for Ungrouped Data

Marks	d $(X - \bar{X})$	Absolute value
17	-14.1	14.1
24	-7.1	7.1
24	-7.1	7.1
24	-7.1	7.1
25	-6.1	6.1
25	-6.1	6.1
27	-4.1	4.1
28	-3.1	3.1
28	-3.1	3.1
30	-1.1	1.1
31	-0.1	0.1
32	0.9	0.9
34	2.9	2.9
34	2.9	2.9
36	4.9	4.9
36	4.9	4.9
40	8.9	8.9
40	8.9	8.9
43	11.9	11.9
44	12.9	12.9
Total		**118.2**

$\sum(d) = 118.2$

Mean deviation = $\frac{\sum(d)}{n} = \frac{\sum(X - \bar{X})}{n} = \frac{118.2}{20} = 5.91$

When these data are grouped as shown in Table 12, calculation of mean deviation is as shown below:

Table 12: A Mean Deviation Table for Grouped Data

Class interval	Class mark (X)	Frequency (F)	F x X	$d(X - \bar{X})$		Absolute value	F (d)
17 – 20	18.5	1	18.5	-12	12		12
21 – 24	22.5	3	67.5	-8	8		24
25 – 28	26.5	5	132.5	-4	4		20
29 – 32	30.5	3	91.5	0	0		0
33 – 36	34.5	4	138.0	4	4		16
37 – 40	38.5	2	77.0	8	8		16
41 - 44	42.5	2	85.0	12	12		24
Total		**20**	**610.0**		**48**		**112**

Mean ($\overline{X}$) = $\frac{\sum(FX)}{\sum F} = \frac{610}{20} = 30.5$

Mean deviation = $\frac{\sum F(d)}{\sum F} = \frac{\sum F\left(X - \overline{X}\right)}{\sum F} = \frac{112}{20} = 5.6$

3.3 Coefficient of Mean Deviation

This is also called relative dispersion and it is usually expressed in percentage. The coefficient of mean deviation is obtained by dividing the mean deviation by the mean and multiplied by 100.

Coefficient of mean deviation = $\frac{\text{Mean deviation}}{\text{Mean}} \text{ x } 100 = \frac{5.6}{30.5} \text{ x } 100 = 18.36$

SELF-ASSESSMENT EXERCISE

From Table 10c find:

i. The range of the distribution.
ii. The mean deviation.
iii. Coefficient of mean deviation.

4.0 CONCLUSION

In this unit, you have learnt the application of range and mean deviation as a measure of dispersion. Other areas discussed here include coefficient of mean deviation. From these discussions, you now know that communication is something we encounter everyday in our lives.

5.0 SUMMARY

The main points in this unit are:

- The range of a set of data is the difference between the largest and smallest number or observation in a set of data
- The mean deviation of a frequency distribution is the mean of the absolute values of the deviation from some measures of central tendency such as mean
- Mean deviation can be calculated from both ungrouped and grouped data
- The coefficient of mean deviation is obtained by dividing the mean deviation by the mean and multiplied by 100.

\

6.0 TUTOR-MARKED ASSIGNMENT

From Table 10 c above, find:

1. The range
2. The mean deviation
3. Coefficient of mean deviation.

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 2 STANDARD DEVIATION AS A MEASURE OF DISPERSION

CONTENTS

1.0 INTRODUCTION

This unit is part of measure of dispersion which explains the application of standard deviation, variance and coefficient of variation. It is hoped that at the end of this unit, you would have learnt and understood the stated objectives.

2.0 OBJECTIVES

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

- define standard deviation
- calculation of standard deviation from both grouped and ungrouped data.
- define and calculate variance from frequency distribution.

3.0 MAIN CONTENT

3.1 Standard Deviation

Standard deviation is the commonly used measure of dispersion of a set of observation. It is also called ‘root mean square deviation’. The name standard deviation implies that it is more standard method of measuring deviations from the mean. For instance, in calculating the mean deviation, the negative sign of individual observation is ignored for convenience. By removing the negative sign is not mathematically justified. A better method of doing it or standardising it is by squaring these deviations so that the negative signs would be eliminated. The

larger the spread of a set of observations, the larger the standard deviation.

Steps in calculating standard deviation

1. Calculate the mean of a set of observations.
2. Find the deviation of each observation from the mean.
3. Square each deviation from the mean.
4. Find the sum of all the squared deviations.
5. Divide the sum of all the squared deviations by the number of observations less one.
6. Divide the square root of the value obtained in 5.

For ungrouped data, the standard deviation is calculated as

$$S = \sqrt{\frac{\sum\left(X_i - \overline{X}\right)^2}{n-1}}$$

Where:

S = Sample standard deviation

X_i = Individual observation

$\overline{X}$ = Mean of the Distribution

N = number of observations

For grouped data, the standard deviation can be

Table 13: A Standard Deviation Table for Ungrouped Data

Marks	$d\ \left(X - \overline{X}\right)$	$d^2\left(X - \overline{X}\right)^2$
17	-14.1	198.81
24	-7.1	50.41
24	-7.1	50.41
24	-7.1	50.41
25	-6.1	37.21
25	-6.1	37.21
27	-4.1	16.81
28	-3.1	9.61
28	-3.1	9.61
30	-1.1	1.21
31	-0.1	0.01
32	0.9	0.81
34	2.9	8.41
34	2.9	8.41
36	4.9	24.01
36	4.9	24.01
40	8.9	79.21
40	8.9	97.21
43	11.9	141.61
44	12.9	166.41

Total	**993.80**

Standard deviation S = $\sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{993.80}{19}} = \sqrt{52.31} = 7.23$

If the data on Table 13 are the whole population, the population standard deviation α is:

$$\alpha = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}$$

Where:

α = Population standard deviation

μ = Population mean

N = Population size

$$\alpha = \sqrt{\frac{993.80}{20}} = \sqrt{49.69} = 7.05$$

For grouped data, sample standard deviation S =

$$\sqrt{\frac{\sum F(X_i - \bar{X})^2}{\sum F - 1}}$$

The data in Table 13 can be grouped as shown in Table 14.

Table 14: Standard Deviation Table for Grouped Data

Class interval	**Class mark (X)**	**Frequency (F)**	**(X - $\bar{X}$)d**	**(X - $\bar{X}$)^{2}d^2**	**F (X - $\bar{X}$)^2**
17 – 20	18.5	1	-12	144	144
21 – 24	22.5	3	-8	64	192
25 – 28	26.5	5	-4	16	80
29 – 32	305	3	0	0	0
33 – 36	34.5	4	4	16	84
37 – 40	38.5	2	8	64	128
41 – 44	42.5	2	12	144	288
Total		**20**		**488**	**916**

$$S = \sqrt{\frac{\sum(Xi - \mu)2}{\sum F - 1}} = \sqrt{\frac{916}{19}} = \sqrt{48.21} = 6.94$$

3.2 The Variance

The variance of a set of data is defined as the square of the standard deviation (S^2). The variance has the same property as the standard deviation. Example, the variance of the above data in Table 14 is $(6.94)^2 = 48.21$.

3.3 Coefficient of Variation (CV)

It is a relative measure of dispersion expressed as percentage. Coefficient of variation is defined as the standard deviation divided by the mean and multiplied by 100.

$$C\,V = \frac{S}{\bar{X}} x\,100$$

Where:

C V = Coefficient of variation
S = Standard deviation
$\bar{X}$ = Mean of the distribution.

Coefficient of variation is used in comparing relative dispersion of two or more sets of observations. The group with higher coefficient is said to have a wider spread.

SELF-ASSESSMENT EXERCISE

The following are the scores of 15 NOUN students in a statistics Examination.
28, 28, 28, 35, 35, 38, 41, 47, 47, 47, 55, 59, 60, 63 and 64.
Find:

a. Standard deviation
b. Group this data into a frequency distribution, using a class width of 4, find
   i. The standard deviation
   ii. The variance and
   iii. Coefficient of variation.

4.0 CONCLUSION

In this unit, you have learnt in detail how standard deviation is calculated from both ungrouped and grouped data. You have also learnt how to calculate the variance from standard deviation as well as coefficient of variation.

5.0 SUMMARY

The main points from this unit include the following:

- Standard deviation is the commonly used measure of dispersion of a set of observation. It is also called ‘root mean square deviation.
- Standard deviation can be calculated from both grouped and ungrouped data.
- The variance of a set of data is defined as the square of then standard deviation (S^2).

- Coefficient of variation is defined as the standard deviation divided by the mean and multiplied by 100.

6.0 TUTOR-MARKED ASSIGNMENT

From table 10 c above, find:

1. The standard deviation
2. The variance, and
3. Coefficient of variation.

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

MODULE 5 POPULATION, SAMPLE AND SAMPLING

Unit 1 Population and Sample
Unit 2 Types of Sampling

UNIT 1 POPULATION AND SAMPLE

CONTENTS

This unit introduces you to sampling from entire population. Population or universe is used to mean the totality of cases in an investigation. A population or universe or sample frame implies all possible observations of a given unit of interest. A sample is a subset of a population which is usually taken from a population using an appropriate technique. After studying this unit, you are expected to have achieved the objectives listed below.

2.0 OBJECTIVES

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

- define what a sample is as well as population
- differentiate between a population and a sample
- explain vividly the reasons for sampling.

3.0 MAIN CONTENT

3.1 Population

The term population or universe is used to mean the totality of cases in an investigation. A population or universe or sample frame implies all possible observations of a given unit of interest. For example, if we want to determine the yield of maize in Oyo State in 2012, the population include all farms in Oyo State in which maize was planted in 2012. If

we want to count the population of NOUN students in 2012, each member of the student in 2012 is a subject of the population. In most cases, taking the entire population for some researches is difficult because the size of the population may be too large for an individual to handle.

3.2 Sample

A sample is a subset of a population which is usually taken from a population using an appropriate technique. Sampling on the other hand, is the procedure or technique of selecting a sample from a population. From the example given above, in determining the yield of maize in Oyo State in 2012, it is very difficult and more costly to assess all farms in Oyo State where maize was planted in 2012. Therefore, sampling can be made on maize farms because the population is too large particularly for an individual to handle. Thus, a sample is taken to represent the population. At the end of the analysis, an inference is made about the population of maize farms in Oyo State in 2012. One important thing to note is that the method of sampling that is a good representative of the entire population should be taken into consideration.

3.3 Reasons for Sampling

According to Asika (2006), the reasons for random sampling are:

1. Sometimes, it is practically impossible to make a complete and comprehensive study of the population because of the nature and pattern of distribution or dispersion of the elements of the population.
2. Since sampling enables us to deal with a part of the population, it is obviously cheaper to study a sample than the entire population.
3. Sampling enables us to be more thorough and affords in better supervision than with a complete coverage of the entire population.
4. It enables us obtain quicker results than does a complete coverage of the population.
5. A study of a few of elements will give researcher sufficient knowledge of what obtains in the entire population of study.

SELF-ASSESSMENT EXERCISE

i. Define the term sampling and population?
ii. What are the rationales for sampling?

4.0 CONCLUSION

This unit has introduced you to sampling and population. From this discussion, you must have learnt the meaning of sample and population as well as reasons for sampling.

5.0 SUMMARY

The main points in this unit are:

- A population or universe or sample frame implies all possible observations of a given unit of interest.
- A sample is a subset of a population which is usually taken from a population using an appropriate technique.
- The reasons for sampling include cheaper to study, better supervision as well as obtaining quicker results.

6.0 TUTOR-MARKED ASSIGNMENT

1. What do you understand by the term 'sampling' and 'population'? 2. Give five tangible reasons sampling is necessary.

7.0 REFERENCES/FURTHER READING

Asika, N. (2006). *Research Methodology in the Behavioural Sciences*. Lagos: Longman Nigeria Plc.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 2 TYPES OF SAMPLING

CONTENTS

1.0 INTRODUCTION

This unit is a continuation on sampling. This unit will make you aware of the different sampling types such as random sampling, systematic sampling, multistage sampling, accidental sampling, cluster sampling, purposive sampling and quota sampling.

2.0 OBJECTIVES

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

- define the various types of sampling
- differentiate one sampling techniques from the other
- use appropriate example to explain each type of sampling.

3.0 MAIN CONTENT

3.1 Random Sampling

Generally, there are two types of sampling; random sampling and non random sampling.

Random sampling: In random sampling, the sampling units are selected according to some law of chance which guarantee every element of the population an equal chance of being selected. By given every element an equal chance, the issue of personal bias is eliminated. Sampling units are the element of a population which need to be selected during sampling. Only if the sample is random that we can assume that the distribution of a sample value follows the pattern of distribution. The two main methods used in random sampling are:

a. **Lottery method**: This method ensures that all the elements have equal chance of being selected. Here, the list of each element of the population is made on separate sheet of paper and they are folded. The lists are thoroughly mixed or shuffled and the

selector can close his eyes before every picking and reshuffle continuously until the desired sample size is obtained.

b. **The use of table of random numbers:** This is through the use of specially prepared table called 'table of random numbers'. In using table of random numbers, all the elements of the population are numbered, depending on the size of the population and the sample size required. The numbering is done in the following pattern:

   i. If the number of elements is less than100, the numbering is between 01 and 99
   ii. If the number of elements is between 100 and 1000, the numbering is between 001 and 999.
   iii. If the number of elements is between 1000 and 10,000, the numbering is between 0001 and 9999 and so on.

Stratified random sampling: This is one type of random sampling in which the population is first divided into two or more groups called strata. Then random selection is made within each stratum. Here, consideration is given to the nature of population elements in the sample frame. This method ensures representativeness. This method is very important where the population is heterogeneous (not similar). For example, if a sample of NOUN academic programme be drawn, it is very important to categorise the respondents into diploma, undergraduate and postgraduate so that the sample will be a good representative of the university. If only undergraduates or postgraduates are used, the report would not be a good representation of the university.

3.2 Non Random Sampling

In non-probability sampling, it is not all the elements of a population have equal chances of being selected. Examples of non- probability sampling are:

i. **Systematic sampling**: This is the one in which the sampling units are selected at fixed interval from the population. When a complete list of the population is available, a common method of selecting a sample is to be select every N_{th} item from the population. For example, if samples of 100 items are to be selected from a population of 1000, the first step is to determine the intervals the items are to be selected. This could be achieved by defining the interval X, where X is the reciprocal of the desired number of samples.

$$X=\frac{1000}{100}=10$$

Then the interval X is 10. Therefore, the first item to be picked could range from 1 to X. example, if the first item is 1, the second item to be picked could be (1 + X) = 11, the third would be (11 + X) = 21 and so on. If the first item chosen is 5, the second item would be (5 + X) = 15. The third item to be selected would be (15 + X) = 25 and so on.

ii. **Multistage sampling**: Here, the sampling is done in stages. The population is distributed into a number of stages. In the first stage, some samples are taken by some peculiar method. From this sample, the sample units are subdivided into second stage and some samples are taken and so on until the desired sample size is achieved. Example, in a socio-economic survey of a country, first stage may involve dividing the country into regions and some states would be selected from each region. The second stage may involve chosen some LGA's from the states selected. The third stage may involve selecting some towns or villages from each LGA and then take the desired sample size from each town or village.

iii. **Cluster or area sampling**: In many situations, the units in the population exist in some natural groups. The population is arbitrarily divided into groups called Clusters. In some cases, where there are many clusters, a random selection of clusters is made and from each cluster selected, the desired sample size is made.

iv. **Accidental or haphazard sampling**: This is a method of sampling in which the respondent is received by accident or chance. It is whoever comes across your way is sampled. Examples, the selection of people in public places to ask questions by journalists

v. **Purposive sampling**: This is a sampling technique used to measure particular characteristics or attribute of a population. In this case, the investigator has definite purpose in mind about the desired sample and specific controls are adopted to achieve the desired objective. For example, if an investigator is interested in socio-economic survey of maize farmers, he will purposively select maize farmers for his investigation. Any other crop farmers will not be involved in the investigation.

vi. **Quota sampling**: It is a sampling method in which a population is divided into different quota based on some agreed percentage. Quota system attempts a fair representation of different classes that may exist in a given population. It is commonly used in

market research surveys as well as public opinion. Selection of sample is guided by a set of quota control such as a set of people to be studied, gender and age or L.G.A.

SELF-ASSESSMENT EXERCISE

Explain the different types of sampling with relevant examples

4.0 CONCLUSION

In this unit you have learnt the different types of sampling. It is good to note that the type of sampling to be employed depend on the nature of survey the researcher is carrying out. Some sample surveyed need to be good representative of the population why some must not necessarily be.

5.0 SUMMARY

The main points in this unit include the following:

- There are two major classes of sampling, random sampling and non random sampling.
- Random sampling involves the selection of elements from a population without bias, all the elements in the population have equal chances of being selected.
- Non random sampling such as random sampling, systematic sampling, multistage sampling, accidental sampling, cluster sampling, purposive sampling and quota sampling.

6.0 TUTOR-MARKED ASSIGNMENT

A politician who has just won an election decided to organise a lunch for his supporters. He can only cater for one hundred and fifty people out of five hundred supporters. What appropriate sampling method would he employ in the selection of his one hundred and fifty supporters from the list of five hundred without bias.

7.0 REFERENCES /FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Asika, N. (2006). *Research Methodology in the Behavioural Sciences*. Lagos: Longman Nigeria Plc, Lagos.

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

MODULE 6 PROBABILITY

Unit 1 Probability and Types of Probability
Unit 2 Rules of Probability

UNIT 1 PROBABILITY AND TYPES OF PROBABILITY

CONTENTS

1.0 INTRODUCTION

In this unit, we are going to learn what probability is all about. It will introduce you to some concepts in probability. Also, the different types of probability are discussed. Probability is achieved by dividing the number of favourable outcomes divided by total number of possible outcomes.

2.0 OBJECTIVES

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

- define what probability is all about
- discuss clearly how probability is applied
- write briefly on the different types of probability.

3.0 MAIN CONTENT

3.1 Probability

The word Probability refers to chance or livelihood. Probability may be based on the use of past experiences in making rational decision. It may also involve some past event about which we have little or insufficient information. Suppose that an event E can happen in S ways out of a total

of n possible equally likely ways, then the probability of occurrence of the event (success) is denoted by P

$$P = Pr(E) = \frac{S}{n}$$

The probability of the event (E) not occurring (failure) is denoted by q

$$q = Pr(\bar{E}) = \frac{n-s}{n} = 1-\frac{S}{n} = 1-P = 1 - Pr(E)$$

Therefore, $P+q=1$ or Pr (E) +Pr (n of E) or ($\bar{E}$) = 1.

Probability is therefore defined as quantification of the livelihood of an event occurring based on past experience. In general terms, probability of event E occurring is defined as

$$Pr(E) = \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

It can also be defined as a relative frequency.

Example, If a die is tossed without bias, what is the Probability of

i. 2 showing up (ii) either 1 or 2 showing up (iii) 1 not showing up.

Solution:

(i) Pr (1) = 1/6

(ii) Pr (1 or 2) = 1/6 + 1/6 = 2/6 = 1/3

(iii) Pr (1) = 1-1/6 = 5/6

.

3.2 Types of Probability

i. **Objective probability**: This is a type of probability of an event that is based
on empirical experience which expresses a universal truth. Such experiment can be repeated to arrive at similar conclusion. For example, in the tosses of a fair coin, the probability of getting head is 0.5.

ii. **Subjective probability**: These are quantifications of personal judgements which may not be based on past experiences. It can be an expression of personal assessment or value judgement which is subjective in nature. For example, one may stipulate that based on personal experience, fat students are more intelligent than slim students.

iii. **Mutually exclusive events:** Two or more events are called mutually exclusive if the occurrence of any of them excludes the occurrence of the others. That is, if one event occurs, the other cannot occur again. Then,
Pr (E_1E_2) = 0. Example, if there are 50 male and 40 female in a class, if a person is picked, the Probability of picking a male =

50/90 =5/9. The Probability of picking a female = 40/90 = 4/9.

iv. **Joint probability**: Two or more events are said to form a joint event, if they can occur at the same time. Example, if there are 15 black balls and 30 white balls in a box, what is the Probability of picking a black ball and a white ball, if the two balls are picked simultaneously from the box?
Solution: Pr (BW) = 15/45 × 30/45 = 1/3 × 2/3 = 2/9.

v. **Conditional probability**: The conditional Probability of an event A is the Probability of the event A occurring given that another event B has occurred. The Probability of an event changes as conditions regarding its sample space change. It is usually written as Pr (A/B), which is probability of A occurring given that B has occurred.
$P(A/B) = \frac{P(AB)}{P(B)}$ Provided $P(B) \neq 0$.

The conditional Probability depends on whether the two events are dependent or independent.

SELF-ASSESSMENT EXERCISE

Describe the various types of probability and explain how it is applied.

4.0 CONCLUSION

This unit has introduced you to probability and the application of probability. The term probability is used by many people not only in statistics as a subject but in our daily expressions in which the deeper meaning is not understood by many people. The different types of probability such as objective probability, subjective probability, joint probability. Mutually exclusive events and conditional probability were also discussed in detail.

5.0 SUMMARY

The main points in this unit include the following:

- Probability is the number of favourable outcomes divided by the total possible outcome.
- Objective probability is a type of probability of an event that is based on empirical experience which expresses a universal truth.
- Subjective probability is the quantifications of personal judgements which may not be based on past experiences.
- Two or more events are called mutually exclusive if the occurrence of any of them excludes the occurrence of the others.
- Two or more events are said to form a joint event, if they can occur at the same time.

6.0 TUTOR-MARKED ASSIGNMENT

Example 2: 100 students of NOUN Undergraduate students were picked at random for a survey, comprising twenty five 100 level students, fifteen 200 level students, twenty 300 level students, twenty two 400 level students and eighteen 500 level students. What is the probability that a student picked is from

(i) 200 level students
(ii) 400 level students
(iii) either 100 level or 500 level students

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 2 DEPENDENT AND INDEPENDENT EVENTS AND RULES OF PROBABILITY

CONTENTS

1.0 INTRODUCTION

This unit is a continuation of probability aspect of the course. It will introduce you to certain concepts such as dependent and independent events. We will look at these concepts in details and see how they are applied. Also, certain rules guiding probability will be treated in detail. The objectives below specify what you are expected to have learnt after studying this unit.

2.0 OBJECTIVES

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

- give some definition of certain concepts in probability
- apply some rules of probability
- differentiate one rule from another.

3.0 MAIN CONTENT

3.1 Dependent and Independent Events

Dependent events: Two or more events are statistically dependent if the occurrence of one affects the occurrence of the other. Under dependent events, the conditional probability of an event A given that event B has occurred is equal to the probability of event A and B occurring simultaneously divided by the probability of event B.

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

$$P(B/A) = \frac{P(AB)}{P(A)}.$$

So, $P(AB) = P(B)\,P(A/B)$

$P(BA) = P(A).\,P(B/A)$.

Example, if there are 20 balls in a box of which 14 are black and 6 are white. Out of the 14 black balls, 8 are perforated and 6 are good. Out of the 6 white balls, 3 are perforated and 3 are good. If a ball is selected at random, what is the probability that it is perforated.

Solution: In all, 11 balls are perforated while 9 are good.
Total possible outcome = 20, number of favourable outcome = 11.
P (E) = Number of favourable outcome = 11
Total possible outcome = 20 = 0.55
By applying this formula of conditional probability of dependent event, the probability of selecting a perforated white ball is
P (PW) = 3/20
The probability of getting any white ball is
P (W) = 6/20
Therefore, the probability of selecting a perforated ball, given a white ball is :

$$P(P/W) = \frac{P(PW)}{P(W)}$$

$$= \frac{11/20 \text{ x } 6/20}{6/20} = \frac{11}{20},$$ which is still the same answer as above.

Independent events*:* Two or more events are statistically independent, if the occurrence of one does not affects the occurrence of the other. In other words, Probability of occurrence of one remains same while the other may occur or not. The Probability of two joint independent events A and B is the product of the two probabilities.
P (AB) = P (A) × P (B)
P (BA) = P (B) × P (A)
For the conditional probability under independent event, we can now substitute P(AB) by P(A) × P(B).
If the conditional probability of event A given that B has occurred is generally given as:

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

But P(AB) = P(A) × P (B)

$$\text{Therefore, } P(A/B) = \frac{P(A) \text{ x } P(B)}{P(B)} = P(A)$$

$$P(B/A) = \frac{P(B) \text{ x } P(A)}{P(A)} = P(B)$$

Example: Given the above example, the probability of selected a perforated ball given a white ball is
P (P/W) = P (P) = 11/20
You will still arrive at the same answer.
Therefore, the conditional probability of two independent events (A & B) is therefore given as
P (A/B) = P (A)

P (B/A) = P (B).

3.2 Rules of Probability

1. If P is the probability that an event will happen (Success) in a single trial, and q is the probability that the event will not happen (failure), then P = 1 - q, or P + q =1.
2. The probability of the occurrence of one or another of two or more mutually exclusive events is the sum of the probabilities of the separate events.
   P (A or B) = P(A) + P(B)
3. The probability of occurrence of two or more independent events is the product of the separate probabilities.
   P (A and B) = P(A) × P(B)

Example 1: If the probability that a machine A will spoil in the next 2 years is 0.6 and the probability that a machine B will spoil in the next 2 years is 0.3. Find the probability that both the machines will spoil in the next two years.

Solution: The Probability that A and B will spoil in the next 2 years = P (A and B)
= P (A) × P (B) = 0.6 × 0.3 = 0.18
Example 3: Assuming 60 balls (green and brown) from three colour categories (deep, ordinary and light) were selected from a box. If 14 (8 green and 6 brown) balls were deep, 36 (22 green and 14 brown) balls were ordinary and 10 (6 green and 4 brown) balls were light. What is the probability of a ball drawn at random being:

i. A green ball (ii) a brown ball (iii) a deep ball (iv) an ordinary ball (v) a light brown and a deep green ball (vi) either deep green or an ordinary brown ball.

Solution: Here, a table is required to simplify the information provided in the question.

Table 15: Colour Categories of 60 Balls

Colour category	Deep	Ordinary	Light	Total
Green	8	22	6	36
Brown	6	14	4	24
Total	14	36	10	60

i. Probability of picking a green ball = P (G) = probability of picking a deep green + probability of picking ordinary green + probability of picking a light green.
   P (G) = P (DG) + P (OG) + P (LG)

= 8/60 + 22/60 + 6/60 = 36/60 = 3/15.

ii. P (B) = P (DB) + P (OB) + P (LB)
= 6/60 + 14/60 + 4/60 = 24/60 = 2/5.

iii. P (D) = P (DG) + P (DB) = 8/60 + 6/60 = 14/60 = 7/30.

iv. P (OB) = P (OG) + P (OB) = 22/60 + 14/60 = 36/60 = 3/5.

v. P (LB and DG) = P (LB) × P (DG)
= 4/60 × 8/60 = 1/15 × 2/15 = 2/225/

vi. P (either DG or OB) = P (DG) + P (OB)
= 8/60 + 14/60 = 22/60 = 11/30.

SELF-ASSESSMENT EXERCISE

A problem in Mathematics is given to two NOUN students A and B whose chances of solving the problem is $\frac{1}{2}$ and $\frac{1}{4}$. What is the probability that they will not solve the problem?

4.0 CONCLUSION

In this unit, you have learnt about dependent and independent events as well as rules of probability. The application of the various rules of probability has been discussed.

5.0 SUMMARY

The main points in this unit include the following:

- Two or more events are statistically dependent if the occurrence of one affects the occurrence of the other.
- Two or more events are statistically independent, if the occurrence of one does not affects the occurrence of the other.
- If P is the probability that an event will happen (Success) in a single trial, and q is the probability that the event will not happen (failure), then P = 1 - q, or P + q =1.
- The probability of occurrence of two or more independent events is the product of the separate probabilities.

6.0 TUTOR-MARKED ASSIGNMENT

In a bag containing 8 brown balls and 6 blue balls, two balls are drawn at random

1. If the balls are drawn one after the other without replacement, find the probability that:
(i) Both balls are brown (ii) both balls are blue (iii) the first ball is brown and the second is blue.

2. Find the probability in each case of the above, if the balls are drawn with replacement.

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

MODULE 7 FACTORIAL, PERMUTATION, COMBINATION AND MATHEMATICAL EXPECTATIONS

UNIT 1 FACTORIAL AND PERMUTATIONS

CONTENTS

1.0 INTRODUCTION

In this unit, we discussed factorial and permutations. Here, we will provide basic explanations and calculations of permutation using the formula as shown in the main content. After studying this unit, you are expected to have achieved the objectives listed below.

2.0 OBJECTIVES

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

- define permutation
- calculate the permutation of n objects taken r at a time.

3.0 MAIN CONTENT

3.1 Factorial

Factorial (n!): This is defined as
n! = n (n-1) (n-2)…. 1
Thus 6! = 6 (6-1) (6-2) (6-3) (6-4) (6-5)
= 6 × 5 × 4 ×3 × 2 × 1 =720.
4! = 4 (4 - 1) (4 - 2) (4 - 3)
= 4 × 3 × 2 × 1 = 24.

3.2 Permutation

A permutation of n different objects taken r at a time is an arrangement of r out of the n objects, with attention given to the order of arrangement. The number of permutation of ordered set of r elements taken from a set of n different elements is represented by n P r, which implies permutations of n items taken r at a time.

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

Example 1: In how many ways can six students be seated if 6 seats are available?

Solution:
6 students can be seated in ${}^{n}Pr = {}^{6}P_{6.}$

$${}^{6}P_{6} = \frac{6!}{(6-6)!} = \frac{6!}{0!} = 720 \text{ ways}$$

Note: 0! = 1.
Example 2: In how many ways can five people be seated if only three seats are available?
Solution:
n = 5 and r = 3
5 people can be seated in ${}^{n}Pr$

$$= {}^{5}P_{3} = \frac{5!}{(5-3)!} = \frac{5 \times 4 \times 3 \times 2}{2} = 60$$

Example 2: In a drawer containing 10 bottles of minerals, if two coke bottles are the same, three Fanta bottles are the same and same five Maltina bottles. In how many ways can the then bottles be arranged in a row in the drawer?
Solution:
n = 10, Coke bottles = 2, Fanta bottles = 3 and Maltina bottles = 5.

Permutation $P = \frac{n!}{C!F!M!} = \frac{10!}{2!3!5!} = 2520$

SELF-ASSESSMENT EXERCISE

In how many ways can 10 people be seated on a bench if only 4 seats are available?

4.0 CONCLUSION

In this unit, you have learnt about factorial and permutation and their applications. A permutation of n different objects taken r at a time is an arrangement of r out of the n objects, with attention given to the order of arrangement.

5.0 SUMMARY

The main points in this unit are as follows:

- Factorial (n!): This is defined as n! = n (n-1) (n-2)…. 1.
- A permutation of n different objects taken r at a time is an arrangement of r out of the n objects, with attention given to the order of arrangement.

6.0 TUTOR-MARKED ASSIGNMENT

In how many ways can eight different coloured books be arranged in a row?

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 2 COMBINATION AND MATHEMATICAL EXPECTATIONS

CONTENTS

1.0 INTRODUCTION

This unit is a continuation of module 7. In this unit, we are going to learn how to apply combinations formula into practical applications. After studying this unit, you are expected to have achieved the objectives listed below

2.0 OBJECTIVES

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

- define the term combinations in statistics
- apply the formula in statistical calculations
- define mathematical expectations
- use appropriate examples of calculations that involve combinations
- differentiate combinations from permutations.

3.0 MAIN CONTENT

3.1 Combinations

A Combination of n different objects taken rate is a selection of r out of the n objects, with no attention given to the order of arrangement. The number of combinations of n objects taken r at a time is ${}^{n}Cr$ or C_{r}^{n}

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

In combinations, no attention is given to the order of arrangement and that is what makes it different from permutation. For example, in how many ways can five people be seated if two seats are available?

Under permutation, the order of arrangement on the two seats are:

A B	B A	C A	D A	E A
A C	B C	C B	D B	E B
A D	B D	C D	D C	E C
A E	B E	C E	D E	E D

In permutation, we have 20 ways because attention is given to order of arrangement. In combination, no attention is given to order of arrangement, which means that, AB = BA, AC = CA, AD = DA, AE = EA and so on. It means that instead of having 20 ways, we will be having 10 by applying the formula. The number of ways are:

A B	B C	C D	D E
A C	B D	C E	
A D	B E		
A E			

By applying the formula, nCr = $\frac{n!}{r!(n-r)!}$

$= \frac{5!}{2!(3)!} = \frac{5 \times 4 \times 3 \times 2}{2 \times 3 \times 2} = 10$

Example 2: In how many ways can 11 men football team be selected from a squad of 15, if attention is not given to positions?

Solution:

The combination of 15 items taken 11 at a time is

$15C11 \quad \frac{15!}{11!(15-11)!} = \frac{15!}{11!4!}$

$= \frac{15x14x13x12}{4x3x2} = 1{,}365.$

3.2 Mathematical Expectations

The expected value of an event is obtained by considering the various values that the variable can take and multiply by their corresponding probabilities.

The expected value of variable X is denoted by E(X) which is its probability times the outcome or value of the variable over a series of trials.

$E(X) = \sum X_i P_i$

Where :

X is the random variable which can take the values of xi, x2…., xn

Pi = Respective probabilities P_1, $P_{2\ldots.}$, P_n

The expected value of an item is also considered as the mean of the item.

Example, when a die is tossed, the expected value of X for the probability distribution is;

$E(X) = \sum X_i P_i = 1 \times 1/6 + 2 \times 1/6 + 3 \times 1/6 + 4 \times 1/6 + 5 \times 1/6 + 6 \times 1/6 = 21/6 = 3.5$

3.3 Laws of Mathematical Expectation

1. The expected value of a constant is the constant itself
   E (C) = C, where C = constant.
2. The expected value of the product of a constant and a random variable is equal to the product of the constant with expected value of the random variable
   E (CX) = CE (X).
3. The expected value of the sum or difference of two random variables is equal to the sum or difference of the expected value of the individual random variables.
   E (X ± Y) = (X) ± E (Y)
4. The expected value of the product of two independent random variables is equal to the product of their individual values.
   E (XY) = E (X) x E (Y).

Example 1: In how many ways can a committee of 6 men be chosen from a committee of 10 men if (a) attention is given to position (b) if position is disregarded?

Solution:

a. If attention is given to position, then it is a permutation

$$nPr = {}^{10}P_{12} = \frac{10!}{(10-6)!} = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2}{4 \times 3 \times 2} = 151200$$

b. If position is disregarded, it is a combination

$$nCr = \frac{10!}{6!(10-6)!} = \frac{10!}{6!4!} = \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2} = 210$$

Example 2: In a game of football, a team can get a reward of ₦4000 by winning with probability 0.7 or make a loss of ₦2000 by making a loss with probability 0.3. what is the expectation?

Solution:

$E(X) = \sum X_i P_i = (4000 \times 0.7) + (-2000 \times 0.3) = 2800 - 600 = 2200$

The expectation = ₦2200.

2. The table below shows the value of daily sales of a woman in five days (₦000) with their respective probabilities.

X	4	6	8	10	12
P (X)	1/8	1/6	3/8	1/4	1/12

a. Find the expectation of E (X)
b. Find E (X^2).

Solution:

a. $E(X) = \sum X_i P_i = 4 \times (1/8) + 6 \times (1/6) + 8 \times (3/8) + 10 \times (1/4) + 12 \times 1/12 =$

$= 0.5 + 1 + 3 + 2.5 + 1 = 8$. In thousands, = ₦8000.

b. E (X^2)

X	4	6	8	10	12
P (X)	1/8	1/6	3/8	1/4	1/12
E (X^2)	0.0156	0.02778	0.1406	0.0625	0.00694

$= 0.0156 + 0.02778 + 0.1406 + 0.0625 + 0.00694 = 0.253$

SELF-ASSESSMENT EXERCISE

In how many ways can a committee of five people be chosen out of nine people?

4.0 CONCLUSION

This unit exposed us to the combinations, and how it is applied. Also, you have learnt how to calculate mathematical expectations and the rules guiding mathematical expectations. From this discussion, you must have learnt the difference between permutations and combinations and how each is applied.

5.0 SUMMARY

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

- define and calculate combinations
- differentiate combinations from permutations
- calculate mathematical expectations
- outline the laws of mathematical expectations.

6.0 TUTOR-MARKED ASSIGNMENT

How many different committees of three men and four women can be formed from eight men and six women?

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

MODULE 8 BINOMIAL, POISSON AND NORMAL DISTRIBUTIONS

UNIT 1 BINOMIAL DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

This unit introduces you to a new topic which has some relationship or application of probability which you have already learnt in your previous module. Binomial distribution applies a combination formula plus the product of the probability of success and failure. Thus after studying this unit, certain things will be required of you. They are listed in the objectives.

2.0 OBJECTIVES

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

- define the term binomial distribution
- apply it practically in the solving some problems.

3.0 MAIN CONTENT

3.1 Binomial Distribution

If P is the probability that an event will happen in any single trial (probability of success) and q is the probability that an event will not happen (probability of failure), then the probability that the event will happen e of exactly X times in n trials is given by

$$P(X) = {}^{n}C_{x}P^{x}q^{n-x} = \frac{n!}{X!(n-X)!}P^{x}\,q^{n-x}$$

Where:

X = 0, 1, 2, ... n,
n = total number of trials
P = probability of success
q = probability of failure.

The binomial distribution is a discrete probability distribution of the number of successes or failure in a sequence of independent trials with only two possible outcomes of probability P success and probability of failure (q or 1-P). Such a trial that yields two possible outcomes is called Bernoulli experiment or Bernoulli trial. For example, in tossing a coin, the number of outcomes (sample space) is 2, H for success and T for failure. In tossing a die, there are six possible outcomes, so it is not a Bernoulli trial. If the outcome of the experiment is categorised such that even numbers is success and odd numbers is failure, then it becomes a Bernoulli experiment.

In Binomial distribution, the formula is the product of combinations formula and the probability of success and failure.

Example 1: What is the probability of getting at least five heads in 6 tosses of a fair coin.

Solution

The probability of getting at least five heads is the probability of getting five heads plus the probability of getting six heads.

The probability of getting a head (success) = $\frac{1}{2}$

The probability of getting a tail (failure) = $\frac{1}{2}$

$P(\geq 5 \text{ heads}) = {}^{6}C_{5}P^{5}q^{1} + {}^{6}C_{6}P^{6}q^{0}$

Probability of getting 5 heads = $${}^{6}C_{5} = \frac{6!}{5!(6-5)!}\left(\frac{1}{2}\right)^{5}\left(\frac{1}{2}\right)^{1}$$

$$= \frac{6 \text{ x } 5 \text{ x } 4 \text{ x } 3 \text{ x } 2}{5 \text{ x } 4 \text{ x } 3 \text{ x } 2}\left(\frac{1}{2}\right)^{5}\left(\frac{1}{2}\right)^{1}$$

$= 6 \text{ x } 0.031 \text{ x } 0.5 = 0.094$

The probability of getting 6 heads is

$${}^{6}C_{6}\, P^{6}q^{0} = \frac{6!}{6!(0)!}P^{6}q^{0}$$

$$= \frac{6!}{6!(0)!}P^{6}q^{0} = 1\left(\frac{1}{2}\right)^{6}\left(\frac{1}{2}\right)^{0}$$

= 0.0156 x 1 = 0.0156

The probability of getting at least five heads = 0.0156 + 0.094 = 0.15

Example 2: The probability of getting exactly two heads in six tosses of a fair coin is

$$C^{6}{}_{2}P^{2}q^{4} = \frac{6!}{2!(4)!}P^{2}\,q^{4} = \frac{6\text{ x }5\text{ x }4\text{ x }3\text{ x }2}{2\text{ x }4\text{ x }3\text{ x }2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{4}$$

= 15 (0.25) (0.0625) = 0.234

In a Binomial distribution;

Mean = $\mu = nP$

Variance = $\sigma^2 = nPq$

Standard deviation = $\sigma = \sqrt{nPq}$

Where:

μ = mean of the binomial distribution

SELF-ASSESSMENT EXERCISE

Find the probability of getting two successes in five tosses of a coin if the probability of getting a success in one trial is 1/3.

4.0 CONCLUSION

In this unit, you have learnt about binomial distribution and how it is applied in solving a particular problem.

5.0 SUMMARY

A summary of the major point in this unit is that:

- The binomial distribution is a discrete probability distribution of the number of successes or failure in a sequence of n independent trials with only two possible outcomes of probability P success and probability of failure (q or 1-P).
- In binomial distribution, the formula is the product of combinations formula and the probability of success and failure.

6.0 TUTOR-MARKED ASSIGNMENT

If the probability that a student will pass his examination is 0.6, find the probability that out of five students that were newly admitted:

1. Two will pass their examination.
2. All will pass their examination, using binomial distribution.

7.0 REFERENCES/FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 2 POISSON DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

This unit will make you aware of some definitions and application of Poisson distribution. The Poisson distribution is similar to Binomial distribution. It is employed where the sample size n is relatively large and where the probability of success P is very small compared to the probability of failure. The objectives below specify what you are expected to learn after going through this unit.

2.0 OBJECTIVES

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

- define and explain Poisson distribution
- outline the conditions under which Poisson distribution is applied
- differentiate it from Binomial distribution.

3.0 MAIN CONTENT

3.1 Poisson Distribution

The discrete probability of Poisson distribution is given by

$$P(x) = \frac{\mu^r e^{-\mu}}{r!}$$

Where:

$\mu = nP$ = mean number of successes

e =Mathematical constant

r = Number of successes

In Bernoulli trials, if n is relatively large (n> 5), the application of Binomial distribution is difficult. Therefore, Poisson Distribution is more appropriate when the sample space n is relatively large (5 > n > 50). Not only that, in Poisson distribution, the n may not be readily known. Also, the probability of the event occurring, P is very small relative to the event not occurring, q.

Though, the probability of Poisson distribution has similarities with the Binomial distribution and describes the probability of events that occur

within some given interval. Both Binomial and Poisson distributions deal with events that are independent of each other and are used for events with discrete values. For Poisson distribution, the probability is based on the mean value, μ

Mean = $\mu = nP$

Variance = $\sigma^2 = \mu$

Standard deviation = $\sigma = \sqrt{\mu}$

Example 1: If the probability that a teacher will come to office is 0.2, determine the probability that in 10 days, the teacher will come exactly three times.

Solution

n = 10 and P = 0.2

$\mu = nP = 10\ (0.2) = 2$

$$P\ (r = 3) = \frac{\mu^r e^{-\mu}}{r!} = \frac{2^3 \text{ x } 2.718^{-2}}{3 \text{ x } 2}$$

$$= \frac{8}{6 \text{ x } 2.718^2} = \frac{8}{44.325} = 0.18$$

Example 2:

If the probability of success in a single trial of an experiment consisting of 20 independent trials is 0.2, find the probability of obtaining:

a. 0, 1, 2, 3, 4, 5, using Poisson distribution
b. At most 2 successes
c. At least 3 successes

Solution

M = mean = nP

= 20 x 0.2 = 4

a. The Poisson distribution of obtaining 0 success is

$$P\ (r = 0) = \frac{4^0 \text{ x } 2.718^{-4}}{0!} = \frac{1}{2.718^4} = 0.018$$

$$P\ (r = 1) = \frac{4^1 \text{ x } 2.718^{-4}}{1!} = \frac{4}{2.718^4} = 0.073$$

$$P\ (r = 2) = \frac{4^2 \text{ x } 2.718^{-4}}{2!} = \frac{16}{2 \text{ x } 2.718^4} = \frac{16}{109.15} = 0.147$$

$$P\ (r = 3) = \frac{4^3 x 2.718^{-4}}{3!} = \frac{64}{6 \text{ x } 2.718^4} = \frac{64}{327.45} = 0.195$$

$$P\ (r = 4) = \frac{4^4 \text{ x } 2.718^{-4}}{4!} = \frac{256}{24 \text{ x } 2.718^4} = \frac{256}{1309.81} = 0.105$$

$$P\ (r = 5) = \frac{4^5 \text{ x } 2.718^{-4}}{5!} = \frac{1024}{120 \text{ x } 2.718^4} = \frac{1024}{6549.06} = 0.156$$

b. At most 2 successes = $P\ (r \leq 2)$

$P(r \leq 2) = P(r = 0) + P(r = 1) + P(r = 2)$
$= 0.018 + 0.073 + 0.147 = 0.239$

c. Probability of at least 3 successes $= P(r \geq 2)$
$P(r \geq 2) = P(r = 3) + P(r = 4) + P(r = 5) = 0.195 + 0.105 + 0.156 = 0.546$

SELF-ASSESSMENT EXERCISE

The probability of producing defective tools in a tractor company is 0.1. What is the probability that in a sample of 20 tools chosen at random, exactly 2 will be defective, using:
a. Poisson distribution

4.0 CONCLUSION

In this unit, you have learnt how Poisson distribution is applied and the conditions that make you to apply it instead of Binomial distribution. You have also learnt that in Poisson distribution, the probability depends on the mean value μ.

5.0 SUMMARY

The main points in this unit include the following:

- The Poisson distribution is more appropriate when the sample space n is relatively large ($5 > n > 50$).
- In Poisson distribution, the n may not be readily known. Also, the probability of the event occurring, P is very small relative to the event not occurring, q.
- Though, the probability of Poisson distribution has similarities with the Binomial distribution and describes the probability of events that occur within some given interval.

6.0 TUTOR-MARKED ASSIGNMENT

Example: Assuming an experiment consist of 50 independent trial in which the constant probability of success in a single trial is $p = 0.04$. Find the probability of obtaining exactly:
1. 0
2. 1
3. 2
4. (iv) 3
5. (v) 4 and
6. (vi) 5 successes using the Poisson distribution.

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 3 NORMAL DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

In this unit, you are going to learn about normal distribution which is a continuous probability distribution. You will also learn some properties

of a normal distribution. After studying this unit, you are expected to have achieved the objectives.

2.0 OBJECTIVES

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

- define and explain what normal distribution is all about
- differentiate normal distribution from other distributions
- outline the properties of normal distribution.

3.0 MAIN CONTENT

3.1 Normal Distribution

Normal distribution, also known as normal curve or Gaussian distribution is one of the most important examples of a continuous probability distribution. Assuming we have a variable X, which has a normal curve with mean μ and standard deviation σ, the variable can be transformed to standard form by defining another variable Z, using the formula:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

X = Individual scores

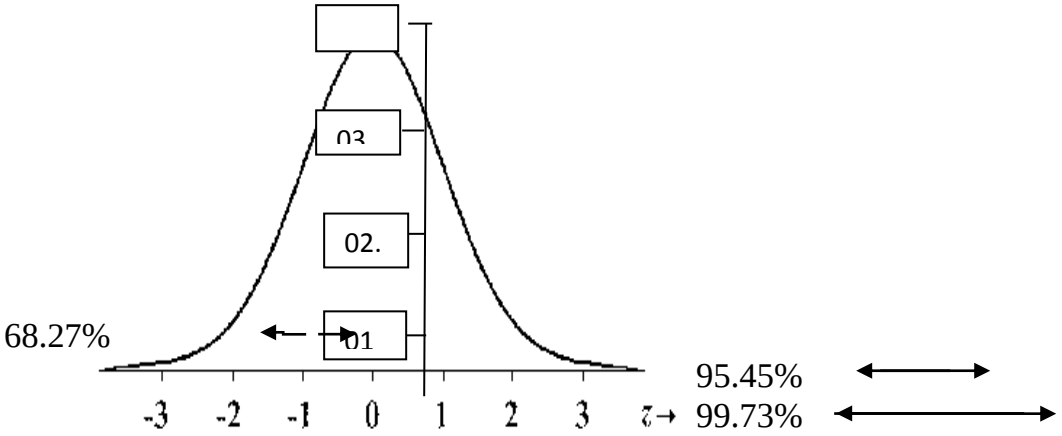
μ = Mean σ = standard deviation

The normal distribution with different means $(\overline{X}s)$ can be made comparable by a means known as standardisation. In the process, individual scores are standardised by subtracting them from the mean of the distribution and divide by standard deviation.

3.2 Properties of a Normal Distribution

i. A normal probability distribution is bell-shaped.
ii. The area within 1 standard deviation of the mean is approximately 068, within 2 standard deviation, the area is approximately 095 and within 3 standard deviation, the area is approximately 0997.
iii. A normal distribution is symmetrical about its mean (μ).
iv. The location and spread of the normal distribution depends on the value of the population mean and the standard deviation.
v. In a normal distribution, the frequency of observation is highest at the mean which is also equal the mode and median.

vi. The range of normal distribution is infinite because the tail never touches the horizontal axis.
vii. When a distribution is normal in shape, we can determine the exact relative standing of any observation in the distribution.
viii. The standard normal distribution is one in which the mean is zero and the variance is 1.



If Z is normally distributed with mean zero and variance 1, it is a standardised normal curve. The areas in the curve shows that Z range from -1 to + 1, -2 to + 2 and -3 to + 3 with their respective total areas of 68.27%, 95.45% and 99.73% respectively.
The standard normal distribution table is used in determining the probability that an observation or score falls within a given interval of the distribution. The area under normal curve is one because the curve is symmetrical. Entries in the table of normal distribution are the areas corresponding to particular values of Z.

$$Z = \frac{\text{Score - Mean}}{\text{Standard deviation}}$$

For example, the area between the mean (μ) and a point Z = 2 standard deviation (2σ)
2σ area = 0.4772
2 (0.4772) = 0.9544, this is why it was stated that the area of the curve between -2 to +2 = 95%.

Example 2: Suppose we want to find the area of Z = 1.64, we check the standard normal distribution table, proceed to the left column of the row Z = 1.6. Then cross to the columns and check under 4. The value = 0.4495.
For Z = 1.96, the value is 0.4750. For Z = 2.32, the value is 0.4898, etc.

Example 3: Suppose that X is a normally distributed random variable with a mean = 8 and standard deviation = 2. Find the probability that X lies in the interval of 6 and 11.

Solution:

Firstly, we find the probability that X lies in the total area of 6

$Z = \frac{X - \mu}{\sigma} = \frac{6 - 8}{2} = \frac{-2}{2} = -1.0 = 0.3413$

Then the probability that X lies in the total area of 11 is:

$Z = \frac{11 - 8}{2} = \frac{3}{2} = 1.5 = 0.4332$

The probability that X lies in the interval 0f 6 and 11 is:

$P(-1.0 < Z < 1.5) = 0.3413 + 0.4332 = 0.7745$

Example 4: If the mean yield of maize is normally distributed with a mean of 360 kg/ha and a standard deviation of 24kg/ha. Find the probability that a farmer chosen at random have:

a. A yield of 420kg/ha
b. A yield greater than 420kg/ha
c. A yield of between 420 and 440kg/ha

Solution

a. Mean = 360kg/ha, σ = 24kg/ha

$Z = \frac{420 - 360}{24} = \frac{60}{24} = 2.5 = 0.494$

b. For a yield of greater than 420kg/ha:

Z = 0.5000 - 0.494 = 0.006

c. For a yield of between 420 and 440, we find the probability of Z = 440kg and add it to the probability of Z = 240kg

$P(Z = 440) = Z = \frac{440 - 360}{24} = \frac{80}{24} = 3.33 = 0.4996$

$P(2.5 < Z < 3.33) = 0.494 + 0.4996 = 0.9936$

SELF-ASSESSMENT EXERCISE

If X is a normally distributed random variable with mean (μ) = 50 and standard deviation (σ) = 10. Find the probability that X will take the value of 60

4.0 CONCLUSION

In this unit, you have learnt about Normal distribution and you have seen how it is applied. Also, you have seen some properties of a normal distribution. If Z is normally distributed with mean zero and variance 1, it is a standardised normal curve. The areas in the curve shows that Z range from -1 to + 1, -2 to + 2 and -3 to + 3 with their respective total areas of 68.27%, 95.45% and 99.73% respectively.

5.0 SUMMARY

The main points in this unit are as follows:

- Normal distribution, also known as normal curve or Gaussian distribution is one of the most important examples of a continuous probability distribution.
- Individual scores of a normal distribution are standardised by subtracting them from the mean of the distribution and divide by standard deviation.
- A normal probability distribution is bell-shaped and symmetrical about it mean.
- When a distribution is normal in shape, we can determine the exact relative standing of any observation in the distribution.
- The standard normal distribution is one in which the mean is zero and the variance is 1.

6.0 TUTOR-MARKED ASSIGNMENT

If the mean yield of a variety of cowpea is normally distributed with a mean of 4800 kg/ha and a standard deviation of 240 kg/ha, calculates the probability that a farmer selected at random have a yield of:

1. 4200 kg/ha
2. Between 4200 and 5000 kg/ha
3. Greater than 4200 kg/ha.

7.0 REFERENCES/FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

MODULE 9 CENTRAL LIMIT THEORY, CONFIDENCE INTERVAL AND HYPOTHESIS TESTING

UNIT 1 CENTRAL LIMIT THEOREM

CONTENTS

1.0 INTRODUCTION

In this unit, you will learn what happens when the sample size of the data you are collecting is getting closer to the population size. Central Limit Theorem describes what happens to the mean, standard deviation and the shape of the mean distribution as the sample size of each of the sample from which the means are calculated becomes larger and larger. The objectives below specify what you are expected to have learnt after studying this unit.

2.0 OBJECTIVES

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

- define and explain what central limit theorem is all about
- use appropriate examples to explain the central limit theory.

3.0 MAIN CONTENT

3.1 Central Limit Theorem

The Central Limit Theorem states that if we take samples of size n from any arbitrary population and calculate the mean, the sampling distribution of mean ($\overline{X}$) will approach the normal distribution as the sample size n increases. Central Limit Theorem describes what happens to the mean, standard deviation and the shape of the mean distribution as the sample size of each of the sample from which the means are calculated becomes larger and larger. The theorem is very central to the probability theory. The mathematical proof is very elaborate. However, the conclusions are very easy to understand. The theorem is considered in three components:

1. The mean of the distribution coincides with the population mean because all the sample means have been reflected in the sample or estimate.
2. The standard deviation of the mean distribution is equal to $\sigma/\sqrt{n}$

   Where:
   n = Sample size
   σ = Standard deviation of the population

This formula indicates two things:

a. When the numerator is increased or decreased, the value of the entire
   expression must be correspondingly increased or decreased.
b. As the denominator is increased, the value of the entire expression must be decreased. This tells us that with a sufficiently large sample, the standard deviation of the mean distribution will be small.
c. As the sample size (n) is increased, the mean distribution tends to have the normal shape, regardless the shape of the population distribution.

In conclusion, the central limit theorem states that the mean distribution tends to be normal as the sample size increases. It has a standard deviation of $\sigma/\sqrt{n}$ and its mean is equal to the true mean of the population from which it is derived. Also, it is central not only in theory but also in the application of our statistical estimations. This implies that many problems can be dealt with by taking samples that is large enough in size and using normal curve to predict the distribution of the sample.

SELF-ASSESSMENT EXERCISE

Explain the central limit theory with relevant examples

4.0 CONCLUSION

In This unit, you have learnt some theories guiding selection of samples from population. The Central Limit Theorem states that if we take samples of size n from any arbitrary population and calculate the mean, the sampling distribution of mean ($\overline{X}$) will approach the normal distribution as the sample size n increases. It states that the mean distribution tends to be normal as the sample size increases

5.0 SUMMARY

The main points in this unit are as follows:

- Central Limit Theorem describes what happens to the mean, standard deviation and the shape of the mean distribution as the sample size of each of the sample from which the means are calculated becomes larger and larger

- The mean of the distribution coincides with the population mean because all the sample means have been reflected in the sample or estimate

- As the sample size (n) is increased, the mean distribution tends to have the normal shape, regardless the shape of the population distribution

6.0 TUTOR-MARKED ASSIGNMENT

The standard deviation of the mean distribution is equal to $\sigma/\sqrt{n}$

Where: n = Sample size and σ = Standard deviation of the population. Briefly explain the interpretation of this formula.

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 2 STATISTICAL ESTIMATION AND CONFIDENCE INTERVAL

CONTENTS

1.0 INTRODUCTION

This unit introduces you to the two types of estimation such as point estimate and interval estimate. It also teaches you to know whether the degree or the percentage confidence or the probability that the actual population parameter lie within a specified range about the sample statistics or not.

2.0 OBJECTIVES

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

- define what statistical estimation is all about
- outline the two methods of estimation
- identify and describe the confidence for estimating the population mean that is contributed with the unbiased estimator of the population mean with sample mean positioned of the estimation interval.

3.0 MAIN CONTENT

3.1 Statistical Estimation

This involves the determination of the value of a known parameter on the basis of a simple statistics. There are two methods of estimation:

1. ***Point estimate***: This is an estimate of a particular parameter given by a single number of the parameter. It is a term used when a simple statistics is used exactly the way it is to estimate a population parameter. For example, the average height of five

sampled men is 36. The 36 is a point estimate because it is a single value. If we say the distance covered by a driver is 50 km, it is a point estimate.

2. ***Interval estimate***: Interval estimate specifies a certain range (interval) of values within which a population parameter is assumed to be. It is an estimate of a population parameter given by two numbers between which the parameters may be considered to lie. Interval estimate are preferred to point estimate because it indicates the precision of an estimate.

Example, we can say the price of one '*mudu'* of maize range between ₦200 - ₦ 250. The distance is 30km + or – 1.30. The age of the household heads range from 25 t- 35 years. Confidence interval is an interval estimation. A statement of the precision or accuracy of an estimate is often called its *reliability.* The method of interval estimation is based on the assumption that the normal probability can be used to compare ranges.

3.2 Confidence Interval

This is the degree or the percentage confidence or the probability that the actual population parameter lie within a specified range about the sample statistics. The confidence for estimating the population mean is contributed with the unbiased estimator of the population mean with sample mean positioned of the estimation interval. The sample mean is positioned at the centre of the centre of the estimation interval. The use of confidence interval enables us to answer the question of how accurate is my estimate? In doing this, the standard error of the mean will acquire a deeper significance. Confidence interval can be determined by using one of the two general formulae, depending on:

i. Whether the population standard deviation is known
ii. Whether the population standard deviation is not known.

Confidence interval (CI) for mean (μ) = $\overline{X} \pm Z\sigma_{\overline{X}}$

Where:

$\sigma_{\overline{X}}$ = Population standard deviation

$CI = \frac{X - \mu}{S}$

Where:

μ = Population mean

S = Standard deviation

X = score or observation.

The value of Z represents the proportion of the area under the normal distribution, thus, the degree of confidence associated with the estimation interval. The most frequently used confidence intervals are 90, 95 and 99% confidence intervals. In social sciences, 95% confidence interval is mostly used. The value of Z to be used in the formula can be gotten from the table of ***standard normal distribution.***

Example: The hourly wages in a particular labour unit are to be normally distributed. A sample of 50 employees' record is examined and the sample mean and standard deviation are 13 and 2.2 respectively. Find the population mean using 95% confidence interval.

Solution

Confidence interval for mean (μ) = $\overline{X} \pm ZS_{\overline{x}}$

Where:

$\overline{X}$ = Sample mean

Z = Standardised formula from Z = $\frac{X-\mu}{S}$

$S_{\overline{x}}$ = Standard deviation of the sample mean

$\sigma_{\overline{x}}$ = Population standard deviation

n= Sample size = 50

We only have population standard deviation, sample standard deviation is not known. To calculate sample standard deviation $(S_{\overline{x}})$, we find the standard error of the mean of the sample $(SE_{\overline{x}})$

$$SE_{\overline{x}} = \frac{S}{\sqrt{n}} = \frac{2.2}{\sqrt{50}} = 0.31$$

At 95% confidence interval, Z = 1.96 (From Z table)

$$\therefore \mu = 13 \pm 0.31 \text{ x } 1.96 = 13 \pm 0.61$$

= 13 + 0.61 = 13.61 Or

13 – 0.61 = 12.39

SELF-ASSESSMENT EXERCISE

What do you understand by the term estimation?

4.0 CONCLUSION

In this unit, you have learnt about the two methods of estimation. Also, confidence interval is the degree or the percentage confidence or the probability that the actual population parameter lie within a specified range about the sample statistics. The value of Z represents the proportion of the area under the normal distribution.

5.0 SUMMARY

A summary of the major point in this unit is that:

- Estimation involves the determination of the value of a known parameter on the basis of a simple statistics
- Point estimate of a particular parameter given by a single number of the parameter
- Interval estimate specifies a certain range (interval) of values within which a population parameter is assumed to be
- Confidence interval is the degree or the percentage confidence or the probability that the actual population parameter lie within a specified range about the sample statistics.

6.0 TUTOR-MARKED ASSIGNMENT

1. Differentiate between point estimate and interval estimate, which one is preferable?
2. How is confidence interval calculated and what percentage is used in social sciences?

7.0 REFERENCES/FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 3 HYPOTHESIS TESTING

CONTENTS

1.0 INTRODUCTION

This unit will introduce you to statistical hypothesis as well as type I and type II errors. Hypotheses are assumptions or guesses or statement about the probability distribution of the population which may be true or may not be true. Thus, after studying this unit, certain things will be required of you. They are listed in the objectives below.

2.0 OBJECTIVES

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

- define what statistical hypothesis is all about
- differentiate between type I and type II errors
- outline the basic steps involved in hypothesis testing.

3.0 MAIN CONTENT

3.1 Statistical Hypothesis

These are assumptions or guesses or statement about the probability distribution of the population. Such assumptions may be true or may not be true. Statistical hypothesis is formulated for the purpose of accepting or nullifying the hypothesis. For example, if we want to decide whether one method of doing things is better than the other, we first establish a hypothesis that there is no difference between the two methods. This type of hypothesis is called null hypothesis (H_o). Direct opposite of null hypothesis is called alternative hypothesis. For example, the alternative hypothesis of the null hypothesis stated above is that there is difference

in the two methods. If one hypothesis P = 0.4, the alternative hypothesis is P ≠ 0.4

3.2 Type I and Type II Errors

When the hypothesis is correct and we reject it, we say that type I error has been made. On the other hand, if we accept the null hypothesis when it should be rejected, type II error is said to have been made. These two types of errors in decision are presented in Table 16.

Table 16: Type of Errors in Decision

Statement	**Decision taken**	
	Null accepted	Null Rejected
Null hypothesis correct	Correct decision	Type I error
Null hypothesis incorrect	Type II error	Correct decision

From Table 16, in either cases of decision, an error has been committed. To minimise errors in decision making, proper procedures have to be followed. In practice, one type of error may be more serious than the other.

The goodness of any decision making procedure, whether it is statistical or otherwise can be measured by the probability of making an incorrect decision. Example, if we are in a court with a judge that has to decide on cases brought to him, he has two options on the suspect, declared guilty or innocent. In this decision, there is unknown truth, guilty or not guilty. This is presented in Table 17.

Table 17: Decision of the Judge for the Unknown Truth

	Unknown truth	
Judges decision	**Guilty**	**Not Guilty**
Guilty	Correct decision	Incorrect decision
Not guilty	Incorrect decision	Correct decision

From Table 17, an error can be made either by convicting innocent person or not convicting a guilty person. Thus, there is the need to protect the rights of the innocent defendant and convict the guilty person.

3.3 Basic Steps Involved in Hypothesis Testing

There are 7 basic steps involved in hypothesis testing:

1. ***Formulate the null and the alternative hypothesis*:** The null hypothesis specifies the parameter value to be tested while the alternative hypothesis specifies the parameter value which is to be tested if the null hypothesis is to be rejected. The null hypothesis (H_o) is to be taken as the principal hypothesis because the decision procedure is carried in relation to the hypothesised value.
2. ***Specify the level of significance to be used*:** This is statistical standard used as basis for rejecting the null hypothesis. For example, if a 5% level of significance (LOS) is specified, the null hypothesis is rejected only if the sample result is so different from the hypothesised parameter value that a difference of such magnitude will occur by chance with a probability of 0.05 or less.
3. ***Select the test statistic*:** The t statistic is used based on the sample used to determine whether the null hypothesis should be rejected or accepted. The t statistics is the sample estimator of the parameter value been tested.
4. ***Establish the critical value of the test statistic*:** Now we determine the critical value of the t statistic. There may be one or more depending on whether it one tail or two tail that is involved. For every type of test, a critical value is in the same unit of measurement as the test statistic and identifies the value of the t statistic that would lead to the rejection of the null hypothesis at the designated level of significance.
5. ***Determine the actual value of the test statistic*:** This is determined through random sample and a collection of data. The sample mean is then calculated.
6. ***Make a decision*:** The obtained value of the sample statistic is compared with the critical value of the test statistic. The null hypothesis is either accepted or rejected.
7. ***Take appropriate managerial action.***

SELF-ASSESSMENT EXERCISE

i. What is a hypothesis?
ii. Differentiate between type I and type II errors.

4.0 CONCLUSION

In this unit, you have learnt about statistical hypothesis. Other areas discussed include type I and type II errors as well as steps involved in hypothesis testing. From this discussion, you have known that when the hypothesis is correct and we reject it, we say that type I error has been made while the reverse is type II error.

5.0 SUMMARY

The main points in this unit are:

- Hypothesis is an assumption or guess or statement about the probability distribution of the population which may be true or may not be true.
- When the hypothesis is correct and we reject it, we say that type I error has been made while the reverse is type II error.
- If we accept the null hypothesis when it should be rejected, type II error is said to have been made.

6.0 TUTOR-MARKED ASSIGNMENT

1. Explain the types of errors that can be committed in statistical hypothesis
2. Briefly explain the basic steps involved in hypothesis testing

7.0 REFERENCES/FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Brink, D. (2010). *Essentials of Statistics*, David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics*. Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

MODULE 10 STUDENT'S t-TEST, Z DISTRIBUTION AND F DISTRIBUTION

Unit 1 Student's t distribution
Unit 2 Z distribution
Unit 3 F distribution

UNIT 1 STUDENT'S t - DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

In this unit you are going to learn how to test hypothesis using student's t test. The student's t test is used instead of Z test when the sample size is relatively small, i e ≤ 30. After going through this unit, you would have achieved the stated objectives of this unit.

2.0 OBJECTIVES

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

- define what student's t -test is all about
- differentiate it from Z test
- use appropriate examples to explain the application of student's t-test.

3.0 MAIN CONTENT

3.1 Student's t– Distribution

If a sample of size n is taken from a normal population, the variable t can then be defined as:

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - \mu\sqrt{n}}{S}$$

Where:

$$S = \text{The sample standard deviation} = \frac{\sqrt{\sum(X - \overline{X})^2}}{n-1}$$

This formula is analogous to the Z statistic given by: $\frac{\overline{X} - \mu}{\sigma/\sqrt{N}} = \frac{\overline{X} - \mu\sqrt{N}}{\sigma}$

The t distribution is usually used for small sample ($n \leq 30$). We find it convenient to replace to replace n with n – 1 and we now have the sample standard deviation ($S_{\bar{x}}$) $S_{\bar{x}} = S/\sqrt{n-1}$. Instead of having Z distribution, we now have t distribution.

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n-1}}$$

Where:

n -1 = the number of degree of freedom (V). The first person that brought about this was W. S. Gosset. He was an employee of Guinness Brewery in Dublin, Island. Guinness Brewery was the first industry to employ statistician and they kept it secret. They did not want to use his name Gosset, rather, they called it student's t distribution after its discoverer who published his work in the name 'student' during the early twentieth century.

The concept of the degree of freedom (V) is used in different areas of statistical analysis. Essentially, the number of degree of freedom indicates the number of values that are free to vary in a random sample. In general, the number of degree of freedom lost is equal to the number of population parameters estimated as the basis of statistical inference. Therefore, the degree of freedom = n -1, where n is the sample size.

Example: Suppose we have a sample that contains five items and we know that the mean is 20, then the sum should be 100. So, we are free to assign arbitrary value to the rest of 4 items while the last one is fixed.
100 – (10 + 25 + 30 + 18) = 83. It means that the 5^{th} item must be 17. So, if we fix the sample mean at $\overline{X}$ in a sample of n, we can assign

values to any variable up to n – 1 item in the sample. Thus, the name degree of freedom = n-1.

3.2. Testing for Differences in Means

In testing for differences in means, there is the need to develop the null hypothesis, i.e. There is no difference in the means. Then the t distribution is given by

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

For $(n_1 + n_2 - 2)$ degrees of freedom, the value of S is obtained as:

$$S = \sqrt{\frac{\sum\left(X_1 - \bar{X}_1\right)^2 + \sum\left(X_2 - \bar{X}_2\right)^2}{n_1 + n_2 - 2}}$$

Example: The monthly sale of a provision seller was 120 cartons. After an increase in the demand for provisions, the mean monthly sale increased to 125 cartons for 20 kiosks and a standard deviation of 15 cartons. Show your views whether the seller is making good sales or not at 5% level of significance.

Solution

Ho = the seller is not making good sale

$\bar{X} = 125$, $\mu = 120$, n = 20 and $\sigma = 15$

$$t = \frac{\bar{X} - \mu}{\sigma/\sqrt{n-1}} = \frac{125 - 120\sqrt{19}}{15} = \frac{5(4.36)}{15}$$

= 1.45

The number of degree of freedom = n – 1 = 19

At 5% LOS, the critical t value = 1.73

Conclusion: The t calculated value is less than the critical value at 5% level of probability. Therefore the H_o is accepted and we conclude that the seller is not making good sale.

SELF-ASSESSMENT EXERCISE

The two largest retail markets in Eastern Nigeria are the markets in the cities of Onitsha and Aba. Comparison of the prices of 10 items in both markets yields the following results:
Onitsha market: mean price of goods = ₦98.90 and standard deviation = ₦83.33

Aba market: mean price of goods ₦90.60 and standard deviation = ₦78.95
At 0.05, is it concluded that goods in Aba market are significantly cheaper than in Onitsha market? (Ayandike, 2009).

4.0 CONCLUSION

This unit has introduced you to the use of student's t-test. It has also shown you how to use student's t-test in the testing for differences in means. The conditions under which the student's t test is applied which is small sample size (≤ 30).

5.0 SUMMARY

The main points in this unit are as follows:

- Student's t- distribution is applied using the formula $t = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - \mu\sqrt{n}}{S}$
- The t distribution is usually used for small sample ($n \leq 30$).
- The number of degrees of freedom = n – 1.
- In testing for differences in means, the formula is $t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
- In this equation above, the degrees of freedom is $(n_1 + n_2 – 2)$

6.0 TUTOR-MARKED ASSIGNMENT

There are two major routes from Kaduna to Lagos, Abuja route and Mando route. A survey was conducted to determine the number of public vehicles that uses the two routes for a period of 10 days. The number of vehicles that uses each route is as follows:

Route A	150	220	240	140	130	210	180	200	170	100

Route B	200	110	220	180	160	240	320	260	110	170

Test at 5% level of probability whether there is significant difference in the number of buses that uses the two routes.

7.0 REFERENCES /FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 2 Z DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

This unit is a continuation of unit 1 which is another method of testing hypothesis. In this unit, you have learnt that Z distribution is applied when sample size is greater than 30. The formula is very similar to that of student's t distribution and it is the same application.

2.0 OBJECTIVES

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

- define and explain the term Z distribution
- outline the conditions under which Z distribution is applied
- differentiate Z distribution from student's t-distribution.

3.0 MAIN CONTENT

3.1 Z Distribution

As a result of central limit theory, the normal distribution (Z distribution) can be used as a basis for determining the critical value for testing a hypothesis. Z is usually employed when $n \geq 30$.

The formula for $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{N}} = \dfrac{\overline{X} - \mu\sqrt{N}}{\sigma}$

In testing for difference in means, the Z distribution can be computed as:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2}}} = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Example: An ADP zonal manager claims that the mean monthly income of VEA is ₦1400. A student wishes to test this claim by contracting 36 VEAs. The monthly income of ₦1400 is given the benefit of doubt. Test whether this claim is true at 5% level of probability, given the following information below:
$H_{o:}$ μ = ₦1400 σ = 50 $\overline{X}$ = ₦1250
Z critical = 1.96

Solution
Ho: $\overline{X} = \mu$ (1250 = 1400)

$$Z = \frac{\overline{X} - \mu}{\sigma} = \frac{1250 - 1400}{50} = \frac{-150}{50} = -3$$

Conclusion: Hence the calculated Z value is less than the critical value (1.96) at 5% level of probability, we accept the null hypothesis that there is no significant difference between the population mean income and the sample mean income.

4.0 CONCLUSION

This unit has introduced you to the application of Z distribution and the conditions under which it is applied. It also teaches you how to make inference about the distribution base d on the calculated value in relation to the critical value.

5.0 SUMMARY

The main points from this unit are as follows:

- Then Z distribution is one of the methods of testing hypothesis.
- The Z distribution is applied when the sample size is relatively large (greater than 30).
- When the Z distribution is calculated, the value is compared with the critical value in order to accept or reject the null hypothesis.

6.0 TUTOR-MARKED ASSIGNMENT

It has been argued that secondary school teachers give birth to more children than university lecturers. A sample of 320 teachers was taken with a mean of 2.95 and standard deviation of 1.98. Also, a sample of 312 university lecturers was taken with a mean of 1.84 and a standard deviation of 0.75. Test this statement at 5% level of significance to see whether this statement is true or not.

7.0 REFERENCES /FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 3 F DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

This unit will make you aware of the application of F distribution and its application. F distribution is used to compare the significant difference in several means at the same time for the analysis of variance (ANOVA). F distribution is used to test differences between three or more means.

2.0 OBJECTIVES

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

- define and explain what f distribution is all about
- differentiate F distribution from other distributions.

3.0 MAIN CONTENT

3.1 F Distribution

The F distribution is used to compare the significant difference in several means at the same time for the analysis of variance (ANOVA). F test is called analysis of variance. ANOVA is a method of reducing to a ratio the proportion of the total variance due to variation between the means of the samples and the proportion due to the variation within the samples as a result of sample error.

F ratio: F ratio is the ratio of the variation between sample means and the variation within the samples.

$$F = \frac{\text{Variation between sample means}}{\text{Variation within the samples}}$$

The degree of freedom from this ratio is used to develop the F distribution curve. There are some similarities and differences between t distribution and F distribution.

Table 18: Differences between t Distribution and F Distribution

	T Distribution	F Distribution
1.	Has one degree of freedom	Has two degrees of freedom, one for numerator and the other for the denominator
2.	Has different t curves for different degrees of freedom	Has different F curves for different pairs of degrees of freedom
3.	t distribution curves are symmetrical	F distribution curves are skewed to the right
4.	t distribution curves is centred at zero	All scores in F distribution are positive
5.	t distribution can be used for testing two means at a time	F distribution can be used for testing several means at a time
6.	t distribution can be used for testing both 1 and 2 tail test of hypothesis	F distribution is used for testing 1 tail test of hypothesis

The name F distribution was named after R. A. Fisher. The F distribution comprises of two degrees of freedom V_1 and V_2. The V1 is the numerator which is = k-1, where K is the number of samples. V2 is the denominator = k (n -1), where n = sample size. n – k if n $\neq$ k.

$$F = \frac{\text{Variation between sample means}}{\text{Variation within the samples}} = \frac{MSA}{MSE}$$

Where:

MSA = Mean square of A

MSE = Mean square due to error

$$MSA = \frac{SSA}{k-1}$$

Where:

$$SSA = \sum_{j=1}^{k} \frac{Tj^2}{n} - \frac{T2}{kn}$$

Where:

Tj = Total for each of the observation in the sample

T = Total observation of all the samples (grand total)

k = number of samples

n = Sample size

$$MSE = \frac{SSE}{k(n-1)}$$

Where:

SSE = Sum square of error

$SSE = \sum_{j=1}^{k}\sum_{j=1}^{n} Xij^2 - \sum_{j=1}^{k}\frac{Tj^2}{n}$, where Xij = summation of the individual observation in all the samples.

Example: In a rabbit feeding experiment, three classes of feed were formulated in three mixture proportions. The following are the result of gains in body weight (kg) of the animals for the three classes:

Mixture	**A**	**B**	**C**
1	2	4	19
2	6	8	20
3	7	6	15

Calculate means of A,B and C, SSA, SSE, MSA, MSE and F

Solution:

Mean of A = $\frac{2+6+7}{3} = 5$

Mean of B = $\frac{4+8+6}{3} = 6$

Mean of C = $\frac{19+20+15}{3} = 18$

We now find the square of individual value for each class

A	A^2	B	B^2	C	C^2
2	4	4	16	19	361
6	36	8	64	20	400
7	49	6	36	15	225
15		18		54	

$Tj^2 = 15^2 + 18^2 + 54^2 = 225 + 324 + 2916 = 3465$

$T = 15 + 18 + 54 = 87$

$T^2 = 87^2 = 7569$

$Xij^2 = A^2 + B^2 + C^2 = 1191$

The null hypothesis is: There is no difference in their means

$H_o = \mu A = \mu B = \mu C$

For the degrees of freedom:

$V_1 = k - 1 = 3 - 1 = 2$

$V_2 = k(n-1) = 3(3-1) = 6$

$F_{(2,6)}^{0.05} = 10.92$

$$F = \frac{MSA}{MSE} = \frac{SSA/k-1}{SSE/k(n-1)}$$

$$SSA = \frac{\sum Tj^2}{n} - \frac{T^2}{kn}$$

$$= \frac{3465}{3} - \frac{7569}{9} = 314$$

$$MSA = \frac{SSA}{k-1} = \frac{314}{2} = 157$$

$$SSE = \sum_{j=1}^{k}\sum_{j=1}^{n} Xij^2 - \sum_{j=1}^{k}\frac{Tj^2}{n}$$

$$= 1191 - \frac{3465}{3} = 36$$

$$MSE = \frac{SSE}{k(n-1)}$$

$$MSE = \frac{36}{6} = 6$$

$$F = \frac{MSA}{MSE} = \frac{157}{6} = 26.16$$

Conclusion: The calculated F value is greater than the critical F value at 5% level of probability. Therefore, the null hypothesis is rejected and we conclude that there is difference in the means of the three samples.

SELF-ASSESSMENT EXERCISE

i. What are the differences between t distribution and F distribution?
ii. Under what condition is each of the distributions applied?

4.0 CONCLUSION

In this unit, you have learnt the application of F distribution and how F distribution is different from t distribution. The name F distribution was named after R. A. Fisher. The F distribution comprises of two degrees of freedom V_1 and V_2. The V1 is the numerator which is = k-1, where K is the number of samples. V2 is the denominator = k (n -1), where n = sample size. n – k if $n \neq k$. F distribution is used to test differences in the means of three or more variables.

5.0 SUMMARY

The main points in this unit are as follows:

- F distribution is used to compare the significant difference in several means at the same time for the analysis of variance (ANOVA).
- F ratio is the ratio of the variation between sample means and the variation within the samples.
- The F distribution comprises of two degrees of freedom V_1 and V_2. The V1 is the numerator which is = k-1, where K is the number of samples. V2 is the denominator = k (n -1).

6.0. TUTOR-MARKED ASSIGNMENT

In one sample of 8 observations, the sum of the squares of deviation of sample values from the sample mean was 84.4 and in the other sample of 10 observations it was 102.6. Test whether the difference is significant at 5% level, given that the five percent point of F for n_1= 7 and n2 = 9 degrees of freedom is 3.29 (Gupta and Gupta, 2010).

7.0 REFERENCES/FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

MODULE 11 CHI SQUARE ANALYSIS

Unit 1 Test of Goodness of Fit (Chi Square Test)
Unit 2 Test of Independence (Chi Square Test)

UNIT 1 TEST OF GOODNESS OF FIT

CONTENTS

1.0 INTRODUCTION

This unit introduces you to chi square (X^2) and its application. The chi square is a non-parametric statistic which is used to determine the difference between expected and the observed results. This unit touches the basic aspect of chi square which is called test of goodness of fit.

2.0 OBJECTIVES

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

- define what chi square is all about
- differentiate chi square distribution from other types of distribution
- apply the formula correctly in any chi square calculation
- outline the properties of a chi square distribution.

3.0 MAIN CONTENT

3.1 Chi Square (X^2) Distribution

The chi square (X^2) test determines whether the observed frequencies of a given observation differ significantly from the frequencies which might be expected from an assumed hypothesis. The chi square is a non-parametric statistic. Generally, the result obtained in a sample does not always agree with expected result. A chi square distribution is used to determine the difference between expected and the observed results. There are basically two methods involved in chi square testing:

i. Test of goodness of fit
ii. Test of independence

3.2 Test of Goodness of Fit

A test of goodness of fit provides a means for testing whether a particular probability such as binomial distribution is close enough approximately to a sample frequency, to the population for which the sample has been drawn. In the test of goodness of fit, the following steps are considered:

1. Establish the null and the alternative hypothesis (Ho and Ha) and select a significant level for the rejection of null hypothesis.
2. Draw a random sample of observation from a population
3. Derive a set of expected or theoretical frequencies under the assumption that the null hypothesis is true.
4. Compare the observed frequencies with the expected or theoretical frequencies by subtracting the expected frequencies from the observed (deviations)
5. Find the square of the differences (deviations) in each observation
6. Divide the square of the differences by the expected or theoretical frequencies
7. Find the sum of step 6
8. If the aggregate difference is too large to be attributed to chance fluctuation at the selected level of significance, the null hypothesis is rejected and the alternative accepted.

3.3 Properties of Chi Square Distribution

1. Chi square (X^2) is skewed to the right.
2. The distribution is concentrated on the right hand side of the curve (no negative value).

3. The exact shape of the distribution depends on the degree of freedom i.e. there are different chi square distribution curve for different degrees of freedom.
4. The distribution tends to shift to the right and become flatter for large values of the degrees of freedom.
5. The degrees of freedom (V) is usually the number of observations (sample size) minus 1 (k – 1) for test of goodness of fit and (r -1) (c -1) for test of independence, where contingency table is used.

Note: r = number of rows and c = number of columns

Example 1: The following results were obtained from tossing a die 300 times.

No. Turned up	1	2	3	4	5	6
Frequency	43	55	39	56	63	43

1. Test the hypothesis that the die was fairly thrown without bias at 5% level of probability.

Solution:

First, the null hypothesis should be stated:

Ho: There is no difference between the observed (O) and the expected (E) frequencies

Ha: At least 2 of the expected frequencies are different from the observed frequency.

For a die to be thrown, the expected probability is $\frac{1}{6}$

Thus, $\mu = np = \frac{1}{6} \times 300 = 50$

The expected frequency in each case =50

$$X^2 = \sum\left[\left(\frac{O-E}{E}\right)^2\right]$$

Where:

X^2 = Chi square

O = Observed frequency of the distribution

E = Expected frequency of the distribution

No. Turned up	**O**	**E**	**O - E**	**$(O-E)^2$**	$\left(\frac{O-E}{E}\right)^2$
1	43	50	-7	49	0.98
2	55	50	5	25	0.50
3	39	50	-11	121	2.42
4	56	50	6	36	0.72
5	63	50	13	169	3.38

6	43	50	-7	49	0.98
Total					**8.98**

Critical $X^2(0.05, 5) = 11.07$
Degree of freedom V = k -1 = 5
$X^{2 =} 8.98$

Conclusion: The calculated (observed) X^2value is less than the X^2critical value (11.07) at 5% level of probability. Therefore, we accept the null hypothesis that there is no difference between the observed frequency and the expected frequency.

SELF-ASSESSMENT EXERCISE

The figures given below are (a) the observed frequencies of a distribution, and (b) the frequencies of the Poisson distribution having the same mean and total frequency as in (a). Apply the chi square test of goodness of fit.

A	305	365	210	80	28	9	3
b	301	361	217	88	26	6	1

4.0 CONCLUSION

This unit has introduced you to the use of chi square distribution and how to apply it to solve some problems. Chi square (X^2) test determines whether the observed frequencies of a given observation differ significantly from the frequencies which might be expected from an assumed hypothesis. It is a non-parametric statistic which is used to test whether there is significant difference between the observed values and the expected.

5.0 SUMMARY

Here are the main points from this unit:

- The chi square (X^2) test determines whether the observed frequencies of a given observation differ significantly from the frequencies which might be expected from an assumed hypothesis.
- A chi square distribution is used to determine the difference between expected and the observed results.

- A test of goodness of fit provides a means for testing whether a particular probability such as binomial distribution is close

enough approximately to a sample frequency, to the population for which the sample has been drawn.

- The degrees of freedom (V) is usually the number of observations (sample size) minus 1 (k – 1) for test of goodness of fit and (r -1) (c -1) for test of independence, where contingency table is used.

6.0 TUTOR-MARKED ASSIGNMENT

Example 2: 1000 people were randomly selected from Lagos State to determine any of the five brands of soft drinks preferred by them. In order not to be biased, the name of the soft drinks were removed and they are labelled A, B, C, D and E. Determine which of the brands is preferred at 5% level of probability assuming all the brands of soft drinks have the same colour. The table of the brands is as shown below:

Brand	**Frequency**
A	210
B	312
C	170
D	85
E	223
Total	**1000**

7.0 REFERENCES/FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Brink, D. (2010). *Essentials of Statistics*, David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 2 TEST OF INDEPENDENCE

CONTENTS

1.0 INTRODUCTION

This unit is a continuation of unit 1. Test of independence involve the use of contingency table. A two way classification table in which the observed frequency occupies rows and columns is called a contingency table. The objectives below specify what you are expected to have learnt after studying this unit

2.0 OBJECTIVES

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

- define and explain the application of test of independence
- differentiate test of independence test of goodness of fit.

3.0 MAIN CONTENT

3.1 Test of Independence

Test of independence involve the use of contingency table. If the observed frequency occupied a single row as in the case of test of goodness of fit, it is called a one way classification table. Since the number of cells = k , it is also called a 1 x k table (1 by k table). A two way classification table in which the observed frequency occupies rows and columns is called a contingency table. It is a type of table which has one method of classification across the columns and the other across the rows. A contingency table having three rows and three columns is called 3 x 3 contingency table. In general, in row x column contingency table, there are r x c cells. For example, if there are three rows and three columns in a contingency table, there are 3 x 3 cells = 9 cells. The test of hypothesis for two way classification table is the same for one way classification, only that in a one way classification, the degree of

freedom V = (R -1) (C- 1) Where R = number of rows, c = number of columns. For 3 x 3 contingency table, V = (3-1) (3 – 1) = 4.

In general, the expected or theoretical frequency for a cell in the ith row and jth column is calculated as:

$$F_{ij} = \frac{\left(\sum \text{row i}\right)\left(\sum \text{column j}\right)}{\text{Grand total}} = \frac{\sum Ri \sum cj}{N}$$

Where:
R_i = Sum of the frequencies in the row i
Cj = Sum of the frequencies in the column j
N = Grand total

Example 1: Three varieties of sorghum were tested to see whether their yields were dependent on the height of the stalk. The following results were obtained from the field trial of the three varieties. Can it be said that the yield is independent of the height of stalk? Test this at 1% level of probability.

Yields	**Height A**	**Height B**	**Height C**
Yield A	6	7	8
Yield B	7	6	10
Yield C	8	9	11

Solution
H_o= Yield is independent of the height of sorghum varieties
H_a= Yield is dependent of the height of sorghum varieties
The degree of freedom V = (3 – 1) (3 – 1) = 4
The contingency table would be made as shown below:

Yields	**Height A**	**Height B**	**Height C**	**Total**
Yield A	6	7	8	21
Yield B	7	6	10	23
Yield C	8	9	11	28
Total	**21**	**22**	**29**	**72**

The expected frequency for each cell of row i column j is:

$$F_{ij} = \frac{\sum Ri \sum cj}{N}$$

$$R_1C_1(\text{Yield A Height A}) = \frac{21 \text{ x } 21}{72} = 6.13$$

$$R_1C_2(\text{Yield A Height B}) = \frac{22 \text{ x } 21}{72} = 6.42$$

$$R_1C_3(\text{Yield A Height C}) = \frac{29 \text{ x } 21}{72} = 8.46$$

$$R_2C_1(\text{Yield B Height A}) = \frac{21 \text{ x } 23}{72} = 6.71$$

R_2C_2 = (Yield B Height B) = $\frac{22 \times 23}{72}$ = 7.03

R_2C_3= (Yield B Height C) = $\frac{29 \times 23}{72}$ = 9.26

R_3C_1= (Yield C Height A) = $\frac{21 \times 28}{72}$ = 8.17

R_3C_2= (Yield C Height B) = $\frac{21 \times 28}{72}$ = 8.56

R_3C_3= (Yield C Height C) = $\frac{29 \times 23}{72}$ = 11.28

These calculated values are the expected or theoretical frequencies for each of the observed frequencies in the contingency table. The chi square table will now be made as follows:

F_O	F_E	$F_O - F_E$	$(F_O - F_E)^2$	$\frac{(F_O - F_E)^2}{F_E}$
6	6.13	-0.13	0.017	0.0028
7	6.42	0.58	0.336	2.1571
8	8.46	-0.46	0.212	0.0251
7	6.71	0.29	0.084	0.0125
6	7.03	-1.03	1.061	0.1509
10	9.26	0.74	0.548	0.0592
8	8.17	-0.17	0.029	0.0035
9	8.56	0.44	0.194	0.0227
11	11.28	-0.28	0.078	0.0069
Total				**2.44**

Calculated X^2 = 2.44
Critical X^2 = (0.01, 4) = 13.3

Conclusion: The calculated X^2 (2.44) is less than the critical value at 1% level of probability. Therefore, the null hypothesis which states that the yield of sorghum is independent of the height of stalk is accepted while the alternative hypothesis is rejected.

SELF-ASSESSMENT EXERCISE

A certain drug is claimed to be effective in curing colds. In an experiment on 164 people with colds, half of them were given the drug and half of them given sugar pills. The patients' reactions to the treatment are recorded in the following table below. Test the hypothesis that the drug is no better than sugar pills for curing colds (Gupta and Gupta, 2010).

.

	Helped	Harmed	No effect
Drug	52	10	20
Sugar pills	44	12	26

4.0 CONCLUSION

In this unit you have learnt about test of independence. Test of independence involve the use of contingency table. The test of hypothesis for two way classification table is the same for one way classification, only that in a one way classification, the degree of freedom V = (R -1) (C- 1).

5.0 SUMMARY

The main points in this unit are as follows:

- Test of independence involve the use of contingency table.
- A two way classification table in which the observed frequency occupies rows and columns is called a contingency table.
- The test of hypothesis for two way classification table is the same for one way classification, only that in a one way classification.
- The degree of freedom V = (R -1) (C- 1).

6.0 TUTOR-MARKED ASSIGNMENT

Example 2: It was suggested that reader preference for different National Newspapers were independent of geographical locations. A survey was taken in which 300 persons randomly chosen from Enugu, Kano and Ibadan were asked to choose their favourite from among the three newspapers. The following results were obtained.

	Newspaper 1	Newspaper II	Newspaper III
Enugu	75	50	175
Kano	120	85	95
Ibadan	105	110	85

Determine whether readers' preference is dependent on geographical location or not at 5% level of probability.

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

MODULE 12 CORRELATION AND REGRESSION ANALYSIS

UNIT 1 CORRELATION ANALYSIS

CONTENTS

1.0 INTRODUCTION

This unit introduces you to correlation analysis. Correlation is a measure of the degree of association between dependent and independent variables. The direction of relationship between two variables can be observed on a scattered diagram. The precise magnitude of correlation between X and Y can be determined using correlation coefficients called Karl Pearson Correlation or Product Moment Correlation (r).

2.0 OBJECTIVES

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

- define what correlation analysis is all about
- interpret the strength and direction of relationship of the correlation
- determine the correlation coefficient
- determine the Spearman's rank correlation.

3.0 MAIN CONTENT

3.1 Correlation Analysis

Correlation measures the strength and direction of relationship between two or more variables. It is a measure of the degree of association between dependent and independent variables. The dependent variable (Y) is the one whose behaviour is been investigated as been influenced by the independent variables (Xs). For example, if X increase as Y increases, there is positive correlation. However, if Y decreases as X increases, there is negative or inverse correlation. If there is no apparent direction of movement between X and Y, then there is no correlation or X and Y are uncorrelated. If all the points on scattered diagram line up directly, then there is perfect correlation between X and Y. Simple correlation involve one dependent and one i dependent variables (two variables) while multiple correlation involves one dependent and two or more independent variables (more than two variables).

The precise magnitude of correlation between X and Y can be determined using correlation coefficients called Karl Pearson Correlation or Product Moment Correlation (r). The value of r ranges from -1 to +1.

If r = +1 implies perfect relationship, i.e. perfect positive correlation
r = -1 implies perfect negative correlation
r = between 0.81 – 0.99 implies high correlation
r = between 0.61 -0.80 implies substantial correlation
r = between 0.41 – 0.60 implies moderate correlation
r = between 0.21 – 0.40 implies low correlation
r = between 0.05 – 0.20 implies negligible correlation.

However, the square of the value of this r is the r^2 value which is called "coefficient of multiple determination" and is defined as:

$$r^2 = 1 - \frac{\text{variations of Y values from the regression line}}{\text{variations of Y values from their own mean}}$$

The r^2 measures the degree of variability of Y that is explained by X in the equation

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} \quad \text{or}$$

$$r = \frac{\sum XY - \sum X \sum \frac{Y}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

Example: Given the following values of X and Y, calculate the correlation coefficient (r)

X	10	15	13	16	19	20	17	8	$\sum X = 118$
									118
Y	9	14	10	17	15	12	10	5	$\sum Y = 92$

Solution

Applying this formula:

$$\frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}}$$

$$\overline{X} = \frac{\sum X}{n} = \frac{118}{8} = 14.75\mathbf{8}$$

$$\overline{Y} = \frac{\sum Y}{n} = \frac{92}{8} = 11.5$$

X	Y	$X - \overline{X}$	$Y - \overline{Y}$	$(X - \overline{X})(Y - \overline{Y})$	$(X - \overline{X})^2$	$(Y - \overline{Y})^2$
10	9	-4.75		11.85	22.56	6.25
15	14	0.25	2.5	0.63	0.06	6.25
13	10	-1.75	-1.5	2.63	3.06	2.25
16	17	1.25	5.5	6.88	1.56	30.25
19	15	4.25	3.5	14.88	18.06	12.25
20	12	5.25	0.5	2.63	27.56	0.25
17	10	2.25	-1.5	-3.38	5.06	2.25
8	5	-6.75	-6.5	43.88	45.88	42.25
Total				**80.03**	**123.48**	**102.00**

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}} = \frac{80.03}{\sqrt{123.48 \times 102}}$$

$$\frac{80.03}{\sqrt{12594.96}} = \frac{80.03}{112.23} = 0.71$$

r = 0.71

3.2 Spearman's Rank Correlation (R)

Another measure of degree of association or relationship is the Spearman's rank correlation. It is also referred to as coefficient of rank correlation. It is used when we need to rank in measurement. As stated earlier, the square of R is the R^2 called coefficient of multiple determination. R^2 measures the degree of variability in Y that is explained by X in the model. In other words, the R^2 measures the percentage of the influence of X on Y. It is computed using the formula:

$$R = 1 - \frac{6\sum D^2}{n(n-1)}$$

Where:

R = Spearman's rank correlation
D = Differences between two ranks
n = Number of pairs of observation.

Example:

Two students were ranked according to their performance in an examination for the 10 courses they were examined as below. Examine the relationship between their ranks using Spearman's rank correlation.

Courses	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Student 1	2	3	7	1	8	5	10	6	9	4
Student 2	1	4	8	2	7	6	9	5	10	3

Solution

Courses	r1	r2	D (r1 – r2)	D^2
A	2	1	1	1
B	3	4	-1	1
C	7	8	-1	1
D	1	2	-1	1
E	8	7	1	1
F	5	6	-1	1
G	10	9	1	1
H	6	5	1	1
I	9	10	-1	1
J	4	3	1	1

$$\sum D^2 = 10$$

$$R = 1 - \frac{6 \times 10}{10(10^2 - 1)} = \frac{60}{10(99)} = \frac{60}{990}$$

$$R = 1 - \frac{6}{99} = 1 - 0.06 = 0.94$$

$$R = 0.94$$

This means there is strong relationship between the performance of the two students.

SELF-ASSESSMENT EXERCISE

i. Given the following data, calculate the correlation coefficient:

X	1	2	3	4	5
Y	3	5	7	9	11

ii. In a football competition, two teams (A and B) had the following ranks from the seven matches they played. Determine the relationship between their ranks.

Team A	Team B
3	5
5	7
1	2
5	3
3	2
6	4
6	5

4.0 CONCLUSION

In this unit, you have learnt the application of correlation analysis. Correlation measures the degree of association between the dependent and independent variables. The precise magnitude of correlation between X and Y can be determined using correlation coefficients called Karl Pearson Correlation or Product Moment Correlation (r). The value of correlation coefficient ranges from -1 to +1.

5.0 SUMMARY

The main points in this unit are as follows:

- Correlation measures the strength and direction of relationship between two or more variables.
- The precise magnitude of correlation between X and Y can be determined using correlation coefficients called Karl Pearson Correlation or Product Moment Correlation
- The value of r ranges from -1 to +1.

6.0 TUTOR-MARKED ASSIGNMENT

Given the following five observations of variables X and Y, calculate the sample correlation coefficient and interpret your result.

X	Y
2	0
4	3
6	1
7	3
4	0
3	2

7.0 REFERENCES/FURTHER READING

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 2 REGRESSION ANALYSIS

CONTENTS

7.0 References/Further Reading

1.0 INTRODUCTION

This unit introduces you to regression analysis which is similar to unit 1. This unit is another measure of strength and direction of relationship. In addition, regression analysis also tells the rate of change of one variable as a result of a unit change in the other. It describes the effect of one variable or more variables called independent variables on a single variable called dependent variable. Regression analysis gives a deeper explanation of the type, nature and direction of relationship than the correlation analysis.

2.0 OBJECTIVES

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

- define what regression analysis is all about
- explain how regression analysis is applied in statistics
- describe briefly the direction, strength and rate of change of one variable in relation to another.

3.0 MAIN CONTENT

3.1 Regression Analysis

Regression analysis describes the nature of relationship between two variables on the basis of direction, strength and rate of change of the variables. It describes the effect of one variable or more variables called independent variables on a single variable called dependent variable. It is important to clearly differentiate between the dependent and independent variables. The distinction is not always obvious and may depend on the nature and character of the variable.

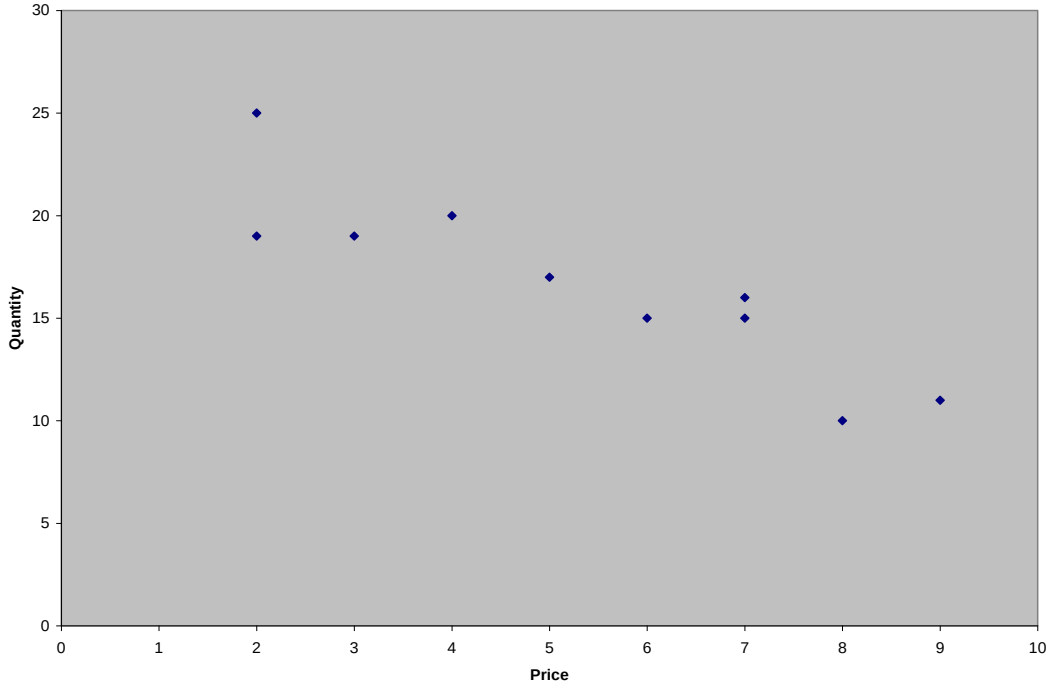
Regression can be classified according to:

1. The number of variables involved, i e simple and multiple regression. Simple regression has 1 dependent and 1 independent variable while multiple regressions have 1 dependent and two or more independent variables. Example of dependent variables is price, income, taste etc.
2. The functional relationship between the dependent and the independent variables, i.e. linear and nonlinear relationship. A scattered diagram can be used to describe the relationship between two variables. For a scattered diagram, dependent

variable (Y) is usually located on the vertical axis while the independent variable (X) is on the horizontal axis.

The Table below shows the record of quantity (kg) of a commodity demanded at various prices (₦)

P	8	7	2	9	4	5	6	2	3	7
Q	10	15	19	11	20	17	15	25	19	16



A scattered diagram showing the record of quantity (kg) of a commodity demanded at various prices (₦).
The statistical relationship between two variables can be estimated by considering these three factors:

i. The direction of relationship
ii. The strength of the relationship
iii. The rate of change of response of one variable to change in the other.

The direction of relationship: The direction of relationship tells us whether the variables are having direct or inverse relationship with one another. The direction of relationship is positive when two variables increase or decrease together. For example, an increase or decrease in one variable will lead to increase or decrease in the other variable respectively. The relationship or direction is negative if an increase in one variable lead to a decrease in the other.

Strength of the relationship: This tells us how closely the points on a scattered diagram about the straight line are. If all the points line up

exactly on a straight line in a scattered diagram, it shows a perfect linear relationship. If the points are scattered randomly with no relationship, it shows no correlation between the two variables.

Rate of change of response: This tells us how much of a change should be expected in one variable if there is a change in the value of the other. This is usually obtained by determining the slope of the regression line.

$$\text{Slope} = \frac{\Delta Y}{\Delta X} = \frac{dY}{dX}$$

SELF-ASSESSMENT EXERCISE

The statistical relationship between two variables can be estimated by considering the three factors, what are they?

4.0 CONCLUSION

In this unit, you have learnt about multiple regression. Multiple regression gives a more detailed analysis of the nature, direction, strength of the relationship between the dependent and the independent variables in a model. In addition to this, it shows the rate of change of response of one variable to change in the other.

5.0 SUMMARY

The main points from this unit are:

- Regression analysis describes the nature of relationship between two variables on the basis of direction, strength and rate of change of the variables.
- The distinction between dependent and independent variables is not always obvious and may depend on the nature and character of the variable.
- The functional relationship between dependent and independent variables may be linear or nonlinear.
- The main types of relationship that multiple regression looks at are the direction of relationship, the strength of the relationship as well as the rate of change of response of one variable to change in the other.

6.0 TUTOR-MARKED ASSIGNMENT

1. Differentiate between a dependent and independent variable.
2. How is regression analysis different from correlation analysis?

7.0 REFERENCES/FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Evans Brothers Nigeria Publishing Limited.

Ayandike, R.N.C. (2009). *Statistical Methods for Social and Environmental Sciences.* Ibadan: Spectrum Books Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics*. Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.

UNIT 3 REGRESSION EQUATIONS

CONTENTS

1.0 INTRODUCTION

This unit is a continuation of the first unit. In this unit, we shall learn about the equations of regression analysis and how it is applied. In this regression, the dependent and the independent variables are usually provided while other parameters such as constant term or the regression coefficient called slope of the relationship will be calculated.

2.0 OBJECTIVES

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

- apply the regression formulae
- determine the constant term a
- determine the regression coefficient b.

3.0 MAIN CONTENT

3.1 Regression Equation

There are two major types of regression equation depending on the nature of the relationship between variables.

Y = a + bX – Linear relationship

Log Y = loga + blog X – Double log

$Y = a + b_1X + b_2X^2$-Quadratic (non Linear)

Where:

Y = Dependent variable

X = Independent variables

a = Constant term

b = slope of the regression line (regression coefficient)

The formula for calculating a and b is given as:

$$b = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} \text{ or } \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - (\sum X)^2}$$

$$a = \frac{\sum Y - b\sum X}{n} \text{ or } \bar{Y} - b\bar{X}$$

When the values of a and b have been calculated, forecasting can be made for values of X which was not observed. Accuracy of the regression line can be measured using the coefficient of determination (R^2).

Example: The data below is the quantity of a commodity supplied at a given prices. Find the value of a and b.

Quantity	2	4	5	6	6
Price	1	2	3	4	5

Solution:

Method 1: Using the first formula:

$$b = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$$

Price (X)	**Quantity (Y)**	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$
1	2	-2	-2.6	5.2	4
2	4	-1	-0.6	0.6	1
3	5	0	-0.4	0	0
4	6	1	1.4	1.4	1
5	6	2	1.4	2.8	4
15	23			10	10

$$\bar{X} = \frac{15}{5} = 3, \ \bar{Y} = \frac{23}{5} = 4.6$$

$$b = \frac{10}{10} = 1$$

$$a = \bar{Y} - b\bar{X}$$

$$= 4.6 - 1\,(3) = 4.6 - 3$$

$$a = 1.6$$

Method 2: Using the second formula:

$$b = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - (\sum X)^2}$$

$$\sum X = 15$$

$$\sum Y = 23$$

X	**Y**	**XY**	**X^2**
1	2	2	1

2	4	8	4
3	5	15	9
4	6	24	16
5	6	30	25
15	23	79	55

$$b = \frac{5 \text{ x } 79 - 345}{5 \text{ x } 55 - 225} = \frac{395 - 345}{275 - 225} =$$

$$= \frac{50}{50} = 1$$

b = 1

$$a = \frac{\sum Y - b\sum X}{n}$$

$$a = \frac{23 - 1(15)}{5} = \frac{8}{5} = 1.6$$

a = 1.6

SELF-ASSESSMENT EXERCISE

The following data gave the rate of fertiliser use and the profit made in raising vegetables on 8 experimental plots. Estimate the regression equation (a and b) and interpret the result.

Y (profit)	9	8	10	3	6	12	14	12
X (Fertiliser)	2	3	6	0	1	7	5	6

4.0 CONCLUSION

This unit has introduced you to the application of regression equation. It has also taught you how to determine the constant term, a and the slope of the relationship (regression coefficient), b.

5.0 SUMMARY

The main points from this unit are as follows:

- Regression equation could be linear or non linear, depending on the nature of relationship that exists between the variables.
- When the values of a and b have been calculated, forecasting can be made for values of X which was not observed.
- Accuracy of the regression line can be measured using the coefficient of determination (R^2).

6.0 TUTOR-MARKED ASSIGNMENT

Consider the following data:

Y	2	7	7	6	8	10	4	5	9	7
X	1	3	2	4	6	6	2	4	5	3

i. Estimate the value of a and b in the regression model using the two formulae.
ii. Predict the value of Y if X = 10

7.0 REFERENCES/FURTHER READING

Afonja, B. (2005). *Introductory Statistics: A Learner's Motivated Approach*. Ibadan: Evans Brothers Nigeria Publishing Limited.

Brink, D. (2010). *Essentials of Statistics,* David Brink and Ventures Publishing, Aps. www.bookboon.com. Accessed 5th June, 2012.

Fernandes, M. (2009). *Statistics for Business and Economics.* Marcelo Fernandes and Ventus Publishing APS, Aps. www.bookboon.com. Accessed 5th June, 2012.

Gupta, C.B. & Gupta, V. (2010). *An Introduction to Statistical Methods* (23rd Revised Edition). New Delhi: VIKAS Publishing House PVT Limited.

Spiegel, M.R. & Stephens, L.J. (1999). *Schaum's Outline of Theories and Problems of Statistics.* USA: McGraw Hill.