STA117

Introduction to Statistics

CODeL

FEDERAL UNIVERSITY OF TECHNOLOGY, MINNA
CENTRE FOR OPEN DISTANCE AND e-LEARNING

FEDERAL UNIVERSITY OF TECHNOLOGY, MINNA
NIGER STATE, NIGERIA

CENTRE FOR OPEN DISTANCE AND e-LEARNING (CODeL)

B.TECH. COMPUTER SCIENCE PROGRAMME

COURSE TITLE

INTRODUCTION TO STATISTICS

COURSE CODE

STA 117

COURSE CODE

STA 117

COURSE UNIT

2

Course Coordinator

D.HAKIMI (Ph.D.)
Department of Mathematics and Statistics
Federal University of Technology (FUT) Minna
Niger State, Nigeria.

Course Development Team

STA 117: INTRODUCTION TO STATISTICS

Subject Matter Experts	Abubakar Usmaan (Ph.D.) FUT Minna, Nigeria.
Course Coordinator	D.HAKIMI (Ph.D.) Department of Mathematics and Statistics FUT Minna, Nigeria.
Instructional System Designers	Oluwole Caleb FALODE (Ph.D.) Bushrah Temitope OJOYE (Mrs.) Centre for Open Distance & e-Learning, FUT Minna, Nigeria
ODL Experts	Amosa Isiaka GAMBARI (Ph.D.) Nicholas E. ESEZOBOR
Language Editors	Chinenye Priscilla UZOCHUKWU (Mrs.) Mubarak Jamiu ALABEDE
Centre Director	Abiodun Musa AIBINU (Ph.D.) Centre for Open Distance & e-Learning FUT Minna, Nigeria.

STA 117 Study Guide

Introduction

STA117: Introduction to Statistics is a 2-credit unit basic course designed for students studying towards acquiring a Bachelor of Technology in Mathematics and Statistics and other related disciplines. The course is divided into 8 modules and 15 study units. It will first take a brief review of the Nature of statistical data and Sources of statistical data. This course will then go ahead to dwell on the method of data collection and preliminary analysis by tables, graphs and charts. The course went further to dwell on the Measures of Central Location: Mean, Mode, Median and weighted means; Measures of dispersion in group and ungroup data, skewness and kurtosis, regression and correlation analysis. Finally, the course dwell on index numbers.

Course Guide

The course guide therefore gives you an overview of what the course; STA117 is all about, the textbooks and other materials to be referenced, what you expect to know in each unit, and how to work through the course material.

What you will learn in this course

The overall aim of this course, STA117 is to introduce you to the basic concepts in statistics and to enable students to have basic knowledge of statistics as it applies to their disciplines.

This course highlights different statistical data and their sources, collection and preliminary analysis by tables and graphs, measures of location and dispersion in group and ungroup data, skewness and kurtosis, simple regression and correlation analysis, and finally, index numbers.

Course Aim

The aim of this course is to introduce students to the statistical processes and methods. It is believed that the knowledge will enable the students to understand the statistical processes and methods in common use. It will be achieved by:

1. Introducing you to fundamental statistical concepts and procedures. Illustrating how these principles can be applied to issues in different fields of endeavour.
2. Explaining to you how data case studies from information-collection agencies and government sources may be employed to solve current problems.
3. Providing the grounding in the ability to generalise and decide positively on situations arising in your discipline.

Course Objectives

It is important to note that each unit has specific objectives. Students should study them carefully before proceeding to subsequent units. Therefore, it may be useful to refer to these objectives in the course of your study of the unit to assess your progress.

You should always look at the unit objectives after completing a unit. In this way, you can be sure that you have done what is required of you by the end of the unit. However, below are overall objectives of this course. On completing this course, you should be able to:

1. Collect adequate and reliable statistical information.
2. Present the data in forms in which the main characteristics are easily understood.
3. Organize and summarize the information collected.
4. Treat the data scientifically i.e. analyze the main features of the data.
5. Deduce meaningfully conclusions or relationships from the statistical analysis.
6. Measure the reliability of the conclusions about a population based on information of the population.

Working through this Course

To complete this course, you are required to study all the units, the recommended textbooks, and other relevant materials. Each unit contains some Self-Assessment Exercises and Tutor Marked Assignment (TMA), and at some points in this course, you are required to submit the tutor marked assignments. There is also a final examination at the end of this course. Stated below are the components of this course to be studied.

Course Materials

The major components of the course are:

1. Course Guide
2. Study Units
3. Text Books
4. Assignment File
5. Presentation Schedule

Study Units

There are 15 study units and 8 modules in this course. They are:

Module One	Unit 1: Nature of Statistical Data Unit 2: Importance of Statistics Unit 3: Sources of Statistics Data Unit 4: Method of Data Collections
Module Two	Unit 1: Data Presentation
Module Three	Unit 1: Measure of Location Unit 2: Weighted Mean
Module Four	Unit 1: Measures of Dispersion I Unit 2: Measure of Dispersion II
Module Five	Unit 1: Skewness and Kurtosis
Module Six	Unit 1: Dependent and Independent Variables Unit 2: Simple Regression

Module Seven	Unit 1: Pearson’s Moment Correlation Unit 2: Spearman’s Rank Order Correlation
Module Eight	Unit 1: Index Numbers

Recommended Texts

These texts will be of enormous benefit to you in learning this course:

Murry R. Spiegel, Outline of Theory and Problems of Statistics (1961),Schaum Publishing Company. U.S.A.

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

Weiss, A.L.(1996), Elementary Statistics, 3rd edition, Addison-Wesley Publishing Co. Inc.

Probability and Statistics, Indira Gandhi National Open University,1991

Schaum’s Outlines Statistics 3rd Edition, Murray R. Spiegel and Larry J. Stephens

Assignment File

The assignment file will be given in due course. In this file, there are details of the work you must submitted to your tutor for marking. The marks you obtain for these assignments will count towards the final mark for the course. Altogether, there are tutor marked assignments for this course.

Presentation Schedule

The presentation schedule included in this course guide provides you with important dates for completion of each tutor marked assignment. You should therefore endeavor to meet the deadlines.

Assessment

There are two aspects to the assessment of this course. First, there are tutor marked assignments; and second, the written examination. Therefore, you are expected to take note of the facts, information and problem solving gathered during the course. The tutor marked assignments must be submitted to your tutor for formal assessment, in accordance to the deadline given. The work submitted will count for 40% of your total course mark.

At the end of the course, you will need to sit for a final written examination. This examination will account for 60% of your total score.

Tutor-Marked Assignment

There are TMAs in this course. You need to submit all the TMAs. The best 10 will therefore be counted. When you have completed each assignment, send them to your tutor as soon as possible and make certain that it gets to your tutor on or before the stipulated deadline. If for any reason you cannot complete your assignment on time, contact your tutor before the assignment is due to discuss the possibility of extension.

Extension will not be granted after the deadline, unless on extraordinary cases.

Final Examination and Grading

The final examination for STA 117 will be of last for a period of 2 hours and have a value of 60% of the total course grade. The examination will consist of questions which reflect the Self-Assessment Questions and tutor marked assignments that you have previously encountered. Furthermore, all areas of the course will be examined. It would be better 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 TMAs and comment on them before the examination. The final examination covers information from all parts of the course.

The Following Are Practical Strategies for Working Through This Course

1. Read the course guide thoroughly

2. Organize a study schedule. Refer to the course overview for more details. Note the time you are expected to spend on each unit and how the assignment relates to the units. Important details, e.g. details of your tutorials and the date of the first day of the semester are available. You need to gather together all this information in one place such as a diary, a wall chart calendar or an organizer. Whatever method you choose, you should decide on and write in your own dates for working on each unit.

3. Once you have created your own study schedule, do everything possible to stick to it. The major reason that students fail is that they get behind with their course works. If you get into difficulties with your schedule, please let your tutor know before it is too late for help.

4. Turn to Unit 1 and read the introduction and the objectives for the unit.

5. Assemble the study materials. Required information needed for a unit is given in the table of content at the beginning of each unit. You will almost always need both the study unit you are working on and one of the materials recommended for further readings, on your desk at the same time.

6. Work through the unit, the content of the unit itself has been arranged to provide a sequence to be followed. As you work through the unit, you will be encouraged to read from your set books

7. Keep in mind that you will learn a lot by doing all your assignments carefully. They have been designed to help you meet the objectives of the course and will assist you pass the examination.

8. Review the objectives of each study unit to confirm attainment.

   If you are not certain about any of the objectives, review the study material and consult your tutor.

9. When you are confident that you have achieved a unit's objectives, you can proceed to the next unit. Proceed unit by unit through the course and try to pace your study so that you can keep yourself on schedule.
10. When you have submitted an assignment to your tutor for marking, do not wait for its return before starting on the next unit. Keep to your schedule. When the assignment is returned, pay particular attention to your tutor's comments, both on the tutor marked assignment form and also written on the assignment. Consult you tutor as soon as possible if you have any questions or problems.
11. After completing the last unit, review the course and prepare yourself for the final examination. Check that you have achieved the unit objectives (listed at the beginning of each unit) and the course objectives (listed in this course guide).

Tutors and Tutorials

There are 8 hours of tutorial provided in Support of this course. You will be notified of the dates, time and location together with the name and phone number of your tutor as soon as you are allocated a tutorial group. Your tutor will mark and comment on your assignments, keep a close watch on your progress and on any difficulties, you might encounter and aid you during the course. You must mail your tutor marked assignment to your tutor well before the due date. At least two working days are required for this purpose. They will be marked by your tutor and returned to you as soon as possible.

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

1. Further clarification is required on any part of the study units or the assigned readings.
2. You have difficulty with the self-test or exercise.
3. You have questions or problems with an assignment, with your tutor's comments on an assignment or with the grading of an assignment.

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

GOOD LUCK!

Table of Content

Module 1

Nature and Sources of Statistical Data

Unit 1

Nature of Statistical Data

Content

1.0 Introduction

2.0 Learning Outcomes

3.0 Learning Content

3.1 Statistical data and methods

3.2 Types of statistics

4.0 Conclusion

5.0 Summary

6.0 Tutor-Marked Assignment

7.0 References/Further Reading

1.0 Introduction

The word statistics is used in a variety of ways. Everything dealing remotely with the collection, processing, analyzing, interpretation and presentation, of numerical/non-numerical data belongs to statistics. These include such tasks as the collection and presentation of data on admission, marriages, revenue etc. The word statistics is often used to mean either numerical description of statistical data or statistical methods.

2.0 Learning Outcomes

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

1. Define statistical data and statistics
2. Conceptualize how data are derived from narrative description/information.

3.0 Learning Content

3.1 Statistical Data and Method

3.1.1 Statistical Data

This refers to numerical description of quantitative aspect of things. This description may take the form of counts or measurement. Therefore, statistics of admission may include the number of students admitted from different schools or from different studies of the federation. It also can mean the data from the different states of the federation.

Example:
Tax remittance from some federal ministries.

Ministry	**Amount (N) Billions**	**Percentage (%)**
Education	3.5	26.92
Industries	1.0	7.69
N. B. S	2.5	19.23
Foreign Affairs	2.0	15.38
Finance	1.9	14.62
Information	2.1	16.15
Total	**13**	**100**

3.1.2 Statistical Method

Statistical method is a technique used to obtain analyses or present numerical data. Examples of statistical techniques include:

i. Collection and assembling of data.

ii. The classification and condensation of data
iii. Presentation of data in either tabular form or graphical form.
iv. Analysis of data.

Self-Assessment Exercise(s) 1

1. What do you understand by the word data in relation to statistics?
2. What is statistical method?

3.2 Types of Statistics

Definition:

Statistics is the science of decision making in the face of uncertainty. We meet uncertainties when we toss a coin, when we experiment with a new drug, when we try to predict the outcome of an election or game. etc.

We can also define statistics to mean a science that deals with the process of Collecting, Organizing, Presenting, analyzing, and utilization of numerical data from a sample to make valid inferences and decisions about a population in economics, business and other social and physical sciences.

There is no doubt that it is virtually impossible to understand a good part of the work done in natural and social sciences without having at least a sparking acquaintance with statistics. Numerical data derived from surveys and experiments constitute the raw material on which interpretation, analysis and decision are based and it is essential to be able to 'squeeze' out usable information from the data. This is in fact the major objective of statistics. Therefore, statistics is defined as a branch of mathematics that deals with the collection, organization, classification, analysis, interpretation, presentation of data expressed in numerical form for the purpose of drawing valid conclusions. This statistics is a theory of information with inferences making as its objective.

Application of statistics can be divided into the following broad areas:

i. Descriptive statistics
ii. Inferential statistics
iii. Inductive statistics

The descriptive statistics is concerned with processing, summarizing conclusions and presentation of data. This could be in the form of tables or graphs to reveal some inherent information contained in a data set; and to present the information in convenient form.

The inferential statistics utilizes sample data to make estimates, decisions, prediction or generalization about a population.

The inductive or analytical statistics is concerned with the method of analysis. Statistics enables us to be able to evaluate data sets intelligently.

Self-Assessment Exercise(s) 2

1. Define Statistics.
2. Explain the following:
   i. Descriptive statistics
   ii. Inferential statistics
   iii. Inductive statistics

4.0 Conclusion

In this unit you should have learnt the concept of data, as the foundation of all statistical analyses and how data are derived from repeated events, information or scientific observation as well as the types of statistics.

5.0 Summary

In this unit, we have defined data as the reduction to numerical figures a body of information and that data are basis of all scientific investigation and analysis. The decision maker who has no data to back his claims is unlikely to convince many listeners. We also identified three types of statistics, namely: descriptive, inferential and inductive statistics.

6.0 Tutor-Marked Assignment

1. What do you understand by the word data in relation to statistics?
2. What is statistical method?
3. State the types of statistics?

7.0 References/Further Reading

Indira Gandhi National Open University, (1991) Probability and Statistics.

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

Unit 2

Importance of Statistics

Content

1.0 Introduction

We can also define statistics to mean a science that deals with the process of Collecting, Organizing, Presenting, analyzing, and utilization of numerical data from a sample to make valid inferences and decisions about a population in economic, business and other social and physical sciences.

2.0 Learning Outcomes

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

1. State the purpose of statistics.
2. Appreciate the importance of statistics.

3.0 Learning Content

3.1 Importance of Statistics

'Statistics' has come to occupy a more prestigious position in the curricula of various 'courses.' What hitherto used to be the problem of Applied sciences has now become universal tool to understand and interpret phenomena. Statistical techniques are now being increasingly used in diversified disciplines such as in Business, Economics, Banking and Military intelligence.

Statistics and Business: There is hardly any area in which the impact of statistics is felt more strongly than in business, where, as a Way of life, decisions which affect profitability and losses must be made at all levels of all kinds of businesses. In the last four decades, the application of statistical methods brought about drastic changes in all the major areas of business management, general management, research development, finance production, sales, advertising and the rest.

Of course, not all problems in these areas are statistical in nature, but the list of those which can be either partly or entirely by statistical methods is very long. Let us now consider some of the roles or importance of statistics in business. Statistics helps businessmen in planning and formulation of future policies.

Manufacturer must know in advance 'how much is to be produced', 'how many workers and how much raw materials will be needed to produce that estimated quantity' and 'what quality, type, size, color or grade of the product is to be manufactured.' In short, he must have a production plan. Such a plan which requires all the details given above cannot be framed without quantitative facts.

Statistical methods of analysis are helpful in the marketing function of an- enterprise through its enormous help in market research, advertisement campaigns and in comparing - the sales and performances.

Statistical methods of analysis provide an important tool, to the - management of- a business enterprise for cost and budgetary control. Various statistical techniques such as index numbers and time series analysis help in the study of price behavior; regression and correlation help in the estimation of relationships between two or more variables e.g. relationships are established between market demands and per capita income, sales promotion and profit enhancement. etc.

The theory and techniques of sampling can be used in connection with various business surveys with a considerable savings in time and money. These techniques are now being extensively used in the test checking of accounts. Statistical quality control is now being used in Industries for establishing quality standards for products, for maintaining the requisite quality, and for assuring that goods sold out are of acceptance given standard.

Statistics and Economics: Statistical data and methods of statistical analysis render valuable assistance in the proper understanding of the economic problems and the formulation of economic policies. Most economic problems always involve facts which can be expressed numerically, e.g. volume of trade, wages, prices, bank deposits, clearing House returns, output of industries, etc.

These numerical magnitudes are the outcome of a multiplicity of causes and are therefore subject variations from time to time. Such economic problems as highlighted above are very suited to statistical treatment. For example, before the ‘vision 2010’ committee can recommend a working policy that will tackle the problems of unemployment in the country, the committee must have knowledge about the following: Is unemployment increasing or decreasing?

Is it widespread or largely confined to certain areas? Does it affect the educated and uneducated like or is more pronounced in any particular class or profession? Which industries are expanding and which are folding up? All these questions can be answered statistically, and the resultant data will enable the committee form a correct policy on unemployment.

Apart from economic policy, the developments of economic theories have also been facilitated by the use of statistics. The increasing importance of statistics in the study of economic problems has resulted in a new branch of study called Econometrics.

Statistics and Science: Statistics is now making a great contribution to the solution of many problems in Medicine, Biology, Agriculture, Meteorology, Physical and Chemical sciences. For example, in the fields of medicine and public health the recently developed and rapidly growing science of biostatistics is applying powerful mathematical and statistical methods to the study of such fundamental problems connected with the growth, development, illnesses, and deaths of human populations as the harmful effects of air pollution, the relationship between diet and heart diseases and between smoking and lung cancer.

In all these areas, statistical methods provide, a framework for looking at the problems in a systematic and logical way, In the biological sciences, statistics is used as an aid to the intelligent planning of experiments, and as a means of assuring the significance of the results of such experiments. Such experiments could be about the growth of animals under different diets and environments

In Agriculture, statistics is very important in experimental design. Experiments such as crop yields with different seeds, fertilizers and types of soil, etc. are frequently designed and analyzed according to statistical principles. In fact, the entire theory of heredity rests on statistical basis, and its development has been intimately related to the development of statistics.

In the physical sciences, statistics is not only applied. in the calculation of standard error, and fittings of curves but also in the treatment of the complex problems of molecular, atomic and nuclear structure. Furthermore, statistical methods have been developed in the fields of astrology, geology, meteorology and physics.

Statistics and Mass Communication: Mass Communication is the dissemination of information or messages through a channel to a target mass audience. These channels can either be through electronic media (television, radio, etc.) or print media (newspapers, magazines, journals, etc.) The agents of this communication system are the Journalists, Broadcasters, Public Relation Practitioners and Advertisers.

Some of the roles of statistics to these agents of communication are highlighted thus statistics enables journalists and reporters to gather their information and tabulate it into the simplest form for easy and better understanding. In this way, information that is lengthy, clumsy and difficult to understand are simplified and made easily comprehensible. When writing on or dealing with numerical data for the understanding of the general public, statistical diagrams and graphs, because of their greater appeal to the eye and imagination tender valuable assistance to journalists and reporters.

The understanding of sampling technique (a statistic4 method) enables the agents of communication (especially. journalists and Audience Research' departments of radio stations) in selecting an unbiased and appropriate sample when carrying out opinion polls, surveys or researches. B example, N.T. A's claim of over 30 million viewers, O.G.B.C' claims of widest coverage in the entire southwest of Nigeria. Also, investigations by Media research and Advertising agencies can only be considered valid when adequate sampling techniques are used.

Statistics also equip the agents of the communication system 5in the area of prediction. Sports journalists and writers use statistical past records and data for predicting sporting activities, particularly in football as evident in poor forecasting It is also important for advertising practitioners to know how to state the results of their surveys or findings numerically. Conclusions stated numerically are definite, more appealing and hence, more convincing to the members of the public than conclusions stated qualitatively. Example, detergent advertisement carries the caption, 'we have so more detergents this year'.

This caption is definitely less attractive than saying.' out of every 10 house wives, 6 prefer our detergent'. Newspaper or magazine editor may apply statistical techniques to simplify and select stories that will occupy the little spaces available. This is usually affected on the layout sheets where the pages are planned for the stories, tables, diagrams, pictures, etc. to fit in.

Lastly, journalists and reporters must be well equipped statistically to be able to understand, write or report effectively on some specialized sectors of the economy such as banking, energy, trades, etc. For instance, only the statistically inclined journalist can understand and hence, write effectively on trading at the Lagos state Stock Exchange market or at the Central Bank Nigeria (CBN) Clearing House.

From the foregoing, it may be concluded that statistics helps us to:

i. Present a large amount of quantitative information
ii. In an organised way, go beyond a meaningless collection of data to a meaningful interpretation
iii. Predict how likely an event will occur
iv. Make inferences from observations
v. Save us time and energy by condensing large amount of information concisely and conveniently in a table.

Self-Assessment Exercise 1

1. What are your purposes for studying statistics?

4.0 Conclusion

In this unit you should have learnt the concept of data, as the foundation of all statistical analyses. You should also have learnt to define statistics, the purpose of statistics and types of statistics.

5.0 Summary

In this unit, we have defined data as the reduction to numerical figures a body of information and that data are basis of all scientific investigation and analysis. The decision maker who has no data to back his claims is unlikely to convince many listeners.

6.0 Tutor-Marked Assignment

1. What are your purposes for studying statistics?

7.0 References/Further Reading

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

Daniel, W.W. and Terrel J. C**.** (1979) Business Statistics: Basic Concepts and Methodology Second Edition, Houghton Mifflin Company Boston.

Unit 3

Sources of Statistical Data

Content

1.0 Introduction

In this unit, we are going to discuss the sources of statistical data as well as the concepts of population and sample. It is necessary to know how fundamental raw material of statistics (data) are derived with emphasize on the relationship between data and statistics.

2.0 Learning Outcomes

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

1. Explain sources of data.
2. Briefly explain the difference between the population and sample.

3.0 Learning Content

3.1 Population and Sample

3.1.1 Population

A population is the set of all the units or objects in a defined area of interest.

For example:

1. All applicants in Niger state.
2. HIV patients in Nigeria.
3. 2007/2008 STA 117 students at FUT, Minna.

In studying a population, there must be a purpose of interest or objective. That is, we may be interested in studying one or more characteristics or properties of the units in the population. This could be the age, height or sex. These characteristics or properties are called variables (because they vary from one unit to the other). Therefore, a variable is defined as a characteristics or property of every unit in the population.
If the population under study is small, it is possible to measure a variable for every unit in the population e.g. The GPA for all 100 level students at FUT, Minna. When we measure a variable for every unit in the population of interest, it is called a CENSUS.

3.1.2 Sample

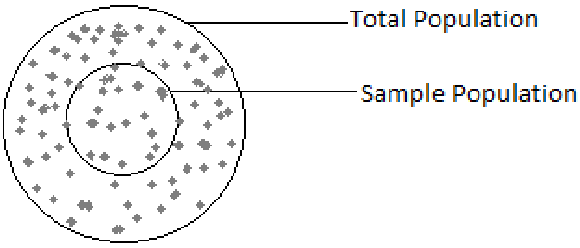
When a portion or subject of the units in the population is selected for study, then, we refer to that portion or subject as a SAMPLE.

Therefore, a sample is a subset or fraction of the units of the population of interest.

Advantages of Sampling:

1. Analysis based on sampling is as precise as that based on the entire population.
2. Use of sampling is time saving.

3. Sampling is cost minimizing-both human and material costs.
4. Analysis based on sample is of greater accuracy than that based on the entire population.
5. The use of population to obtain some of its parameters may not be feasible (practicable) especially with infinite population (population whose number is too large to be known) or when the observation process is destructive. E.g. testing the effectiveness or efficiency of a new vaccine or a new raw material in production.



Self-Assessment Exercise(s) 1

1. What is a population?
2. What is sample?
3. Mention five advantages of sampling.

3.2 Sources of data

After taken a decision on what type of data to be collected, either qualitative, you will the need to collect appropriate data to solve the problem. There are four major sources of statistical data. These are:

1. **Published Sources:** The sources of secondary data include publications of the National Bureau of statistics (NBS), formerly Federal Office of Statistics (FOS). Central Bank of Nigeria (CBN), National population commission (NPC). United Nations (UN), World Health Organizations (WHO) Ministries & Parastatals; others include Journals, Newspapers, Textbooks, and statistical digest. Any data collected from through any of these sources is referred to as published or secondary data
2. **Designed Experiments:** This involves planning a strict control over the units under study which may be people or objects
3. **Sample Surreys:** These are used when information is drawn from a scientifically representative sample of the population of interest. The information so collected is then generalized to cover the entire population.
4. **Direct Observation.**

Self-Assessment Exercise 2

1. State major sources of data.

4.0 Conclusion

In this unit, you should have learnt that statisticians make inferences from representative sample. It is the task of the statistician to find an appropriate method to draw its sample from the population. Also, various sources of data were identified and explained.

5.0 Summary

In this unit, we have explained the concept of population and sample, advantages of sampling. Finally, in this unit, we unfolded sources of data such as published source, designed of experiment, sample survey and direct observation.

6.0 Tutor-Marked Assignment

1. What is a population?
2. What is sample?
3. Mention five advantages of sampling.
4. State four major sources of data collection.

7.0 References/Further Reading

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

Ary and Donald and Jacobs, L.C. (1976) Introduction to Statistics: Purposes and Procedures. New York.

Unit 4

Methods of Data Collection

Content

1.0 Introduction

Statistics' has come to occupy a more prestigious position in the curricula of various 'courses' What hitherto used to be the problem of Applied sciences has now become universal tool to understand and interpret phenomena. Statistical techniques are now being increasingly used in diversified disciplines such as in Business, Economics, Banking and Military intelligence.

2.0 Learning Outcomes

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

1. Discuss characteristics, types and forms of statistical data .
2. Explain methods of data collection

3.0 Learning Content

3.1 Characteristics, Types and Forms of Statistical Data

3.1.1 Characteristics of Statistical Data

1. They must in be aggregate.
2. They must be enumerated to a marked extent by a multiplicity of causes.
3. They must be enumerated or estimated according to reasonable standard of accuracy.
4. They have been collected in a systematic manner for a predetermined purpose.
5. They must be comparable.

3.1.2 Types of Data

The two types of data are:

1. Primary data and
2. Secondary data

Primary data refers to the statistical data (or information) which the investigator originates himself for the purpose of the enquiry at hand.

Secondary data refers to those statistical data which are not originated by the investigator himself, but which he obtained from someone else's records or from some organizations, either in published or recorded forms. There are basically two forms of data.

3.1.3 Forms of Data

A. Quantitative data: These are the data that can be measured on a naturally occurring numerical scale. Like height, weight, volume, examination score. Quantitative data can further be divided into internal and ratio scale. For the ratio scale, the origin (i.e. value zero or o) is a meaningful number where as it is meaningless to interval data. Arithmetic operation, such as additions and subtraction can be performed on ratio scale where as such operations cannot be performed on interval scale.

Examples of quantitative data:

1. Temperature — interval
2. Exam scores — interval
3. Unemployment rate — ratio
4. Convicted cases — ratio.

B. Qualitative Data: These are data that cannot be measured on a natural or numerical scale but can only be classified into categories. Qualitative data can be subdivided into nominal or ordinal data. The ordinal data can be ranked or meaningfully ordered while the nominal data cannot be ordered.

Examples of qualitative data are:

1. Types of fish — nominal
2. Home Town — nominal
3. Colour of dress— nominal
4. Rank in the army — ordinal.
5. Model of cars —ordinal.
6. Health condition — ordinal.

Arbitrary numerical values are assigned to qualitative data for ease of computer entry and analysis. These assigned values are simply codes which cannot be added, subtracted, multiplied or divided.

Examples are:

1- No Decision; 2-Strongly Disagreed; 3-Disagree;

4-Agree; 5-Strongly Agree. Or 1- Success; 0 – Failure.

Self-Assessment Exercise(s) 1

1. Enumerate five characteristics of data.
2. What are the two types of data we have?
3. Name two forms of data and their examples.

3.2 Methods of Data Collection

For any statistics survey/census, the data collection stage is an important activity for testing the success or the failure. If serious deficiencies occur which are not caught early and rectified, there is little that can be done to achieve the ultimate goal of enhancing good quality data. Hence, it is therefore necessary to ensure careful and thorough planning in the organization of field work.

Generally, enumeration phase involves many different activities such as organization of the field structure, recruitment and training of field staff, design of questionnaire, determination of data collection procedures, and quality control over the field operations, among others. These can be divided under three (3) broad topics namely; Data collection methods, Data collection plans and Data collection approaches.

1. Data collection plans: As part of the collection plans, the under listed steps must be taken to ensure the desired level of quantity:
2. Good frame — A frame is a list or map of the universe or population to be covered.
3. Development of instruments including pre-testing, Questionnaires and manuals.
4. Field organization/Recruitment
5. Budgeting
6. Training
7. Publicity/preliminary field arrangement
8. Organization of field work
9. Quality control
10. Time frame
11. Cognizance of problems, especially transportation.
12. Data collection approaches:

    The activity of the data collection could be through:

    i. Censuses — this is the approach used when information is required on every member of the population of interest.
    ii. Micro- study: - Micro-study is a term that covers field investigations undertaken on small or unique scale. The basic reason for undertaking micro-study is to secure in-depth, detailed information on a smaller number of population units.
    iii. Record keeping: - these are systems established for the purpose of gathering fats and commentary about subjects of interest. A record keeping approach could either be formal or informal.

        Enumeration procedures represent the means whereby that data required for either surveys or census purposes are collected. There are various techniques and procedures that may be considered for data collection purposes. However, the key factors to be considered in deciding the enumeration procedures will be the followings:

        i. Types of information needed
        ii. The level of detail required
        iii. Type of analysis
        iv. Stage/state of the development of the country
        v. Nature and dispersion of the population.

Generally, the choices are more limited in developing countries as a result of lower literacy and the absence of communication facilities.

Methods of data collections are:

1. *Direct / personal (face- to —face) interview:* The direct interview method is the most common in this method, enumerators or interviews follow a rigid procedure and asks question already prepared for the purpose. Enumerators are trained in order to standardize the system of data collection in the following components:

i. The coverage and definitions of terms and concepts
ii. How to conduct the interview
iii. How to fill the questionnaire
iv. How to measure certain characteristics
v. How to detect internal consistencies in the replies.

In this method, enumeration should be completed in a short time if adequate staff of enumerators can be trained. The efficiency of this method is dependent on the following factors:

i. Cooperative attitudes of the respondents which will affect the quality and of the data collected.
ii. Type of training provided for field workers
iii. The category I experiment of field workers
iv. Field supervision arrangement.

This method is mostly used in developing countries because of the level of development.

Advantages:

1. In highly complex field or surveys involving multi- subject undertakings, this technique is the most feasible, even where literacy is high.
2. It yields high respondent rate because skilful interviewers can persuade an unwiring respondent.
3. The interviewer is free and has more opportunities to restructure the questions wherever it is necessary to do so.
4. It allows more and accurate information to be obtained by asking the respondent more questions or asking him / her for further explanation.
5. Skilled interviewers will know when to make calls and recalls if respondents are out.
6. It can be applied to all respondents whether literate or illiterate.
7. Interviewer can note the reactions of the respondents if necessary.

Disadvantages:

1. It is very expensive it the sample to the covered is widely scattered geographically.
2. Unscrupulous interviewers may introduce bias by influencing respondent's answers or record to please himself.
3. Respondent bias may be introduced in order to boost his image.
4. Individuals in highly placed positions or high-income group are always difficult to reach.
5. Calls and recalls can increase the cost of the survey.
6. Respondent can give inaccurate or false information due to memory.

7. There is the possibility of misunderstanding or interpretations of questions.

2. Postal or Mail Questionnaire:

This is another technique used in more developed countries and in more literate populations of developing nations. It involves the completion of surrey questionnaires by the respondents themselves. The questionnaires can be distributed by mail (postal questionnaires) or picked up field workers. Generally, it is advisable to restrict the length of self-enumeration questionnaires in order to avoid confusion and reduce non-response. This technique could be referred to as postal or mail questionnaire.

Advantages:

1. It requires smaller field costs even if the sample covers a geographically wide area.
2. In theory, it allows for consultations with other members of the family. Thus, increases the accuracy of the survey data.
3. It eliminates interviewer's bias.
4. It eliminates the problem of interviewing very difficult respondents.

Disadvantages:

5. Lower cooperation and completion rates and less consistent responses.
6. The scope of the survey is usually limited as a result of usage of the technique.
7. Volume of editing and coding at the other data processing stage is usually greater than for other data collection methods.
8. Replies to vital questions may be deliberately avoided or omitted.
9. Answers to vital questions may be ambiguous.
10. It can be used only by literate populations

3. Telephone Interviews:

In some highly developed countries various kinds of data may readily be collected by this method. This method is not popular in developing countries because of the ineffectiveness of the telephone system. Sweden, for example uses this method for their surveys. In this method of data collection, the enumerators telephone the respondents, asking them the desired or need questions for the survey, and the respondents give their answer through the phone.

Advantages:

1. The method, generally, is inexpensive as result of the removed if travel expenses of the advantages of the direct interview.
2. Control and monitoring of interviews are easier.
3. The process of collecting information is faster.
4. It is more flexible.
5. Recalls are faster and cheaper.
6. The response rate is high.
7. It is the method used in accessing very difficult respondent.
8. It reduces embarrassment to respondents.

Disadvantages:

1. It is not a good method where observation is required in an inquiry.
2. The respondent can easily terminate the interview if annoyed by some questions.
3. The method is limited to those with telephones.
4. Cost consideration may reduce the number of questions asked or the time given to the respondent to answer.

4. Direct observation and measurement:

This technique using direct observation and measurements are sometimes used in complicated surveys. An example of observation is a food consumption survey, where the interviewer visits restaurants or homes every day and records consumption on the basis of questioning the respondent and observing what kind of food has been prepared. An example of measurement is where survey respondents are examined by health officials in the course of health — related surveys.

Anthropometric measurement (height, weight and volume) is another survey example or measurements.

Advantages:

1. It gives more reliable and advantages common to direct observation.
2. Useful in undertaking complex surveys.

Disadvantages:

1. It is exceptionally costly
2. It requires highly trained personal
3. Close scrutiny of respondents may influence respondent’s behaviour and responses and findings may be inaccurate.

5. Transcription from records

This is a system or method whereby the needed data/information already exists somewhere in records e. g. marriages, divorce, vital statistics such as records of birth, death, etc.

Advantages

1. Cheap
2. Fast
3. Easy to apply

Disadvantages

1. Not reliable: The data or information gotten from here is as accurate as the original data itself.
2. Official data are not easily made available
3. Combination of Techniques.

Sometimes, combinations of the various data collection techniques are frequently found. For example, an expenditure survey may start with a face-to face interview to obtain background information and to inquire about certain large outlays, where upon the household is asked to keep chary or record of current expenditure for a period of time. An example is the national consumer survey of NBS. Also, a mail inquiry will nearly always require some follow — up by personal visits or telephone for those who do not mail back the completed questionnaires. NBS establishment-based survey uses this technique.

Advantages:
1. It yields high response rate
2. It gives more accurate and reliable information

Disadvantages:
1. It is time Consuming
2. It is very expensive
3. Complex in application and analysis.

Questionnaires

1. A questionnaire contains a sequence of questions relevant to the data or information being sought. Either in census or survey, a questionnaire is needed. Questions in a questionnaire may make or mar data regardless of the qualities of other factors.

Questionnaires are usually of two parts. Part one is the classification section. It contains such details of the respondents like sex, age, marital status, occupation, state of origin, etc. The second part is related to the subject matters of the enquiry.

Characteristics of a Good Questionnaire

1. Questionnaires should be simple and easily understood. Thus, a question on marital status can be put in the following manner.
   MARITAL STATUS (mark (^) in the appropriate answer/box)

   Widowed □
   Single □
   Married □
   Divorced □

2. Questions should be short since very few people enjoy
   form-filling or answering questions.
3. Questions should be able to have precise answer ‘yes’ or ’no’, a date, a number, a place, etc.
4. Questions should not require calculations to be made.
   e.g. Asking the question “what percentage of your annual
   salary is your housing allowance?
5. Questions should not offend, frighten or be teleguiding e.g. don’t know anything about the theft, do you?” is frightening and teleguiding’.

"Is your married life happy?" may offend the respondent if he is
not happily married.

6. Questions that may arouse the questions are not carefully
   worded and may arouse the resentment resentment of the informants should be avoided.
   E.g. questions like "have you stopped smoking"? Or "How often
   do you beat your wife?" these of the respondent.
7. Questions should be of logical order
   e.g. (a) how many bottles of beer do you drink in a day?
   (b) do you drink?
   A logical sequence/order will be:
   (a) do you drink?
   (b) If yes, how many bottles do you drink in a day?

Sampling Frame

A sampling frame contains the basic details of all members of the population from which samples are to be -drawn. Statisticians believe that without a complete sampling frame, a truly random sample cannot be selected. Voters register which contains the names of eligible voters is an example of sampling frame.

Sampling Design

1. In statistics, a sample design is a definite plan, completely determined before any data is collected, for obtaining a sample from a given or known population Some of the most important kinds of sample designs are:
   i. Simple random sampling
   ii. Systematic sampling
   iii. Stratified sampling
   iv. Multi-stage sampling
   v. Quota sampling
   vi. Cluster sampling

Simple Random Sampling

This is a sampling procedure in which every member of the population has equal chance of being selected as a member of the sample. It is mostly (of same kind) population. The methods of selection; include; casting of lots, tossing. of coin or rolling die. All the above methods are however, not perfectly objective. The most objective and widely used method is the random Number table. This is a table; consisting of randomly allocated numbers that are arranged in rows and columns of a standard statistical table.

Systematic Sampling

This is similar to simple random sampling but not exactly the same. It is used when the population is homogenous and fairly small and the sampling frame is complete. An element of randomness is introduced by selecting the first number, member or unit by a random method. The other members are automatically and, systematically selected. The real danger in systematic sampling lies in the possible presence of hidden periodicities. For instance, if it is known that there are usually, mis-prints. In the daily production of a newspaper and this is to be investigated. If we inspect, every 50th copy made by a particular machine; our results would be biased if, because of a regularly recurring failure, every 30th copy has mis–print. .

Method of Selection

We take the population size as N and the sample size as n.

Two cases are possible.

(a) N is divisible by n e.g. $If \quad N = 20 \; and \; n = 4.$

(b) N is not divisible by n e.g. If $N = 28 \; and \; n = 5.$

STEP 1: We assign randomly to every member of the population the units I to N.

STEP 2: We' select, randomly the first member of the sample, say x, such that

$$x < k, where\ K = \frac{N}{n} if Nis\ divisible\ by\ n, or\ K\ is\ taken\ as\ the\ nearest\ integer\ (whole\ number)\ to\ ; \frac{N}{n}\ if\ N\ is\ not\ divisible\ by\ n.e.g\ If\ N = 20\ and\ n = 4\ then\ K = \frac{20}{4} = 5.$$

$$If\ N = 28\ and\ n = 5, then\ K = \frac{28}{5} = 5.6$$

$$\approx 6$$

STEP 3: We select the numbers

$$x, x + k, x + 2k, \dots, x + (n - 1)k.$$

STEP 4: We now take the members of the population originally assigned the numbers in, step 3 '[p;/]'

['p;as our sample.

Stratified Sampling: This is a sampling procedure which consists of stratifying (dividing the population into a number of non —overlapping sub population called strata), then testing a sampling from every stratum. The items or sample from each stratum can then be selected by any suitable random method. Stratified sampling procedure is very good and appropriate when our population is large and heterogeneous. It means that the entire population can be divided into economic sub-groups or class called strata. Although the process of stratification is relatively simple, several substantial problems immediately arise. What should be the basis for stratification? How many strata should be formed? What sample size should be allocated to the different strata? How should the sample within the strata be selected?

Stratification does not guarantee good results, but if successful and properly executed, a stratified sample will generally lead to a higher degree of precision or reliability than a sample random sample of the same size drawn from the population.

In stratification, very often, it is necessary in the selection of the required sample to ensure that the sizes of the respective strata. This is called proportional allocation. In general, if we divide a population of size N into K strata of sizes N_1, N_2 ..... N_k and take a sample of size n_1 from N_1, n_2 from N_2,….n_k from N_k.

We say that allocation is proportional if these ratios are as nearly equal as possible.

Example:
A stratified sample of size n = 60 is be taken from a population of size N = 4000, which consists of three strata of size N_1 = 2000; N_2 = 1,200; N_3 = 800. If sample is to be proportional to size of strata, how many samples should be taken from each stratum?

Solution:

Given n=60,
N_1=2000
N_2=1,200
N_3=800

N=4000

Suffices to find n_1, n_2, n_3.
By definition,

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

$$n_1 = \frac{N_1 \times n}{N}$$

$$= \frac{2000 \times 60}{4000}$$

$$= 30$$

$$n_2 = \frac{N_2 \times n}{N} = \frac{200 \times 60}{4000}$$

= 18

$$n_3 = \frac{N_3 \times n}{N} = \frac{800 \times 60}{4000}$$

= 12.

Check: $n_1 + n_2 + n_3 = n$;

i.e. 30 +18 +12 = 60.

Multi-Stage sampling:

As the name connotes, this sampling procedure include more than one stage. The first stage consists of breaking down the population into sets of district groups. From the sets a number of groups are selected. Each group selected is broken down into units from which a sample is taken. If we stop at this point, we have a two stage sampling. But further stages may be added and the number of stages involved is denoted in the name of the sampling. For instance, five stage sampling means that five stages are involved.

Quota sampling:

In stratified sample, the cost of taking random samples from the individual strata is often so expensive that interviewers are simply given quotas to be filled from the different strata, with very few, (if any) restrictions on how they are to be filled.

This kind of sampling is called quota sampling. For instance, in determining students attitude toward increment in schools fees, in a particular institution, on interviewed can be told to interview, say, 12 students from development A, 5 from department B, 10 from department C , etc; with the actual selection of those to be interviewed being left to the interviewer's discretion.

A major drawback of this method of sampling is that in the absence of firm restrictions on their choice. Interviewers naturally tend to select individuals who are most readily available in order to reduce cost. Thus, in the example given above, a male student of department A conducting the interview may just take his 12 sample in department A from amongst his male friends in the same department. Thus, neglect female students from his department and the entire students in the other departments. Its major advantages are that, it is quick, convenient and relatively cheap to use especially in making research and opinion polls.

Cluster Sampling:

Any sampling procedure usually requires compete sampling frame. Unfortunately, this frame is not always available. Cluster sampling is therefore developed to take care of this inadequacy. In this kind of sampling, the total population is divided into a number of relatively smaller sub- divisions, which are themselves still clusters of smaller units, and then, some of these sub- division or clusters are randomly selected for inclusion in the overall sample.

If the clusters are geographic subdivisions, this kind of sampling is called area sampling. To illustrate cluster sampling, suppose the management of an organization with over 100 branches spread over all the major cities in Nigeria- wants to interview a sample of the workers to determine their attitudes towards a proposed retrenchment exercise if random samples are used to select, say, 10 branches from the list and if some or all employees of these 10 branches are interviewed, the resulting sample, is a cluster sample. This shows that cluster sampling saves time and also minimizes cost.

Self-Assessment Exercise(s) 2

1. State four methods of data collection.
2. Explain what is meant by stratified sampling.

4.0 Conclusion

In this unit, you should have learnt about the characteristics of data, types and forms of data. Different sources of data such as primary and secondary sources as well as stratified sampling, multistage sampling, and cluster sampling as sampling techniques in data collection are treated.

5.0 Summary

In this unit, we have explained features of statistical data and classified data as primary data and secondary data. We have also defined qualitative and quantitative data as forms of data. Finally, in this unit, we explained sources of data such as published source, designed of experiment, sample survey and direct observation.

6.0 Tutor-Marked Assignment

1. Enumerate five characteristics of data.
2. What are the two types of data we have?
3. Name two forms of data and their examples.
4. What do you understand by the following terms?
   i. Questionnaires
   ii. Stratified Sampling
   iii Cluster Sampling
   iv Quotas Sampling
   v. Sampling design
5. State four methods of data collection.
6. Distinguished between Quotas Sampling and Multi-Stage Sampling.

7.0 References/Further Reading

Knapp, R.G (1985). Basic Statistics for Nurses, Second Edition, New York: Delmar Publishers Inc.

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

Module 2

Data Presentation

Unit 1

Data Presentation

Content

1.0 Introduction

This unit is concerned with the meaningful manner in which raw data (which are usually in the form of large set of unorganized numerical values) are summarized and interpreted so that important features and trends may be identified.

A set of data may be presented in tables or described by means of diagrams, charts and graphs. Before discussing these terms, let us look at what you should learn in this unit as stated in the objectives.

2.0 Learning Outcomes

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

1. Tabulate and organize raw data.
2. Present data by means of tables, diagrams, charts and graphs.
3. Interpret and highlight important features of the data through the tables, diagrams, charts and graphs.

3.0 Learning Content

3.1 Narration/Tabulation

After collecting the necessary data, the first task of **a** statistician or researcher is to reduce and simplify the detail into such **a** form that the salient features may be brought out, which will facilitate the interpretation of the assembled data. This procedure is known in classifying and tabulating the data. Tabulation thus enhances the condensation and easy comparison of data. Most published data usually come in tabulated form and it is one of the most popular methods of making data more comprehensible.

	Lagos	**Outside Lagos**	**Total**
Passed	24,375	45,000	69,375
Referred	4,875	22,500	27,375
Absent	5,775	9,000	14,775
Result withheld	1,650	33750	35,400
Failed	825	2,250	3,075
Total	**37,500**	**112,500**	**150,000**

Examples:

1. Guardian Newspapers has three titles in her stables: The Guardian, “African Guardian”, and “Express.” A study of staff ratio in three departments was carried out and the following information was gathered. There are 200 staff in the three

departments, of which 65 are in the African Guardian. Of the staff in the guardian, 30 are in the editorial department and 21 in the advertising department. In the Guardian Express, 22 staff are in the production department and 15 in the advertisement department of the total of 61. The total number of staff in the three departments who worked in the advert department 55 and those in the production department 65.

i. Tabulate the above information so as to give the highest possible information.

ii. How many staff are in the advertisement department in the African Guardian?

Solution:

	The Guardian	**African Guardian**	**Guardian Express**	**Total**
Editorial	30	26	24	80
Production	23	20	22	65
Advert.	21	19	15	55
Total	**74**	**65**	**61**	**200**

1. 19 staff are in the advert department in the African Guardian.
2. In 1998, 150,000 candidates entered for JAMB examinations. 25% of the candidates came from Lagos while the rest came from outside Lagos. Of those who came from Lagos, 65% passed the examinations, 13% were referred'; and of the rest, 0.7% were absent, 0.2% never received their results while the others failed the examinations. Of those who came from outside Lagos, 0.4% passed the examinations, 0.2% were referred and 0.3% had their results withheld. Of the rest, 80% were absent while all others failed the examination.

    i. Arrange the above information on a table.

    ii. What percentage of the candidates failed the examinations?

    iii. Mention two pictorial forms in which the information can be presented.

Solution

ii) Percentage of those who failed

$$\frac{3075 \times 100}{150000}$$

$$= \quad 2.05\%$$

iii) Multiple bar chart and component bar charts.

Self-Assessment Exercise

1. Tabulation thus enhances the condensation and easy comparison of data (Yes or No)?

3.2 Pictorial Presentations (Diagrams, Charts and Graphs)

No matter how informative and well designed a statistical table is, as a medium for conveying to the reader an immediate and clear impression of its content, it is inferior to a good chart or graph. Many people are incapable of comprehending large masses of information presented in tabular form; the figures merely confuse them. Furthermore, many of such people are unwilling to make the effort to grasp the meaning of such data. Graphs and charts come into their own as a means of conveying information in easily comprehensible form. It is for such reasons that government and multinationals always produce popular versions of important white papers in the form of multi-coloured booklets full of simple diagrams and charts. Such diagrams and charts are also often now seen on television both for viewers' easy understanding and for advertising.

Though such pictorial presentation reduces the amount of detail that can be put across to the reader or viewer, often it is not the detail that matters, but rather the overall picture. The most popular charts, diagrams and graphs are:

i. Pie-charts
ii. Bar diagrams (bar charts and histograms)
iii. Graphs (frequency polygons and Ogives)

3.2.1 Pie Charts

A pie-chart is simply a circle divided into sections. This circle represents the total of the data being presented and each section is drawn proportional to its relative size. The main advantage of a pie-chart is that it is easy to understand. It is most suited to very simple comparisons where there are only few groups, say 2 to 4. The use of the pie-chart where there are more than 4 sections to be labelled usually results in a loss of the clear visual effect.

Self-Assessment Exercise 2

1. An investigation of the marital status of the staff of an institution reveals the following:

Marital Status	No. of Staff
Singles	35
Married	130
Widowed	25
Divorced	10

i. Draw a Pie Chart using the above Information.

3.2.2 Bar Charts

A simple bar chart comprises of a number of equally spaced rectangles. A multiple bar chart is usually used in the comparison of two or more attributes.

A component bar chart comprises of bars which are subdivided into components. Example:

Self-Assessment Exercise 3

1. Represent the data in the above SAE 1 in a Bar Chart.

Self-Assessment Exercise 4

The sex distribution of staff in five departments of a Television station is given below:

Department	**Male**	**Female**	**Total**
Admin (I)	25	15	**40**
Programmes (II)	65	30	**95**
Commercial (III)	45	40	**85**
News (IV)	35	15	**50**
Sports (V)	30	10	**40**
Total	**200**	**110**	**310**

Represent the above information on:

i. Multiple bar chart

ii. Component bar chart

3.2.3 Histograms

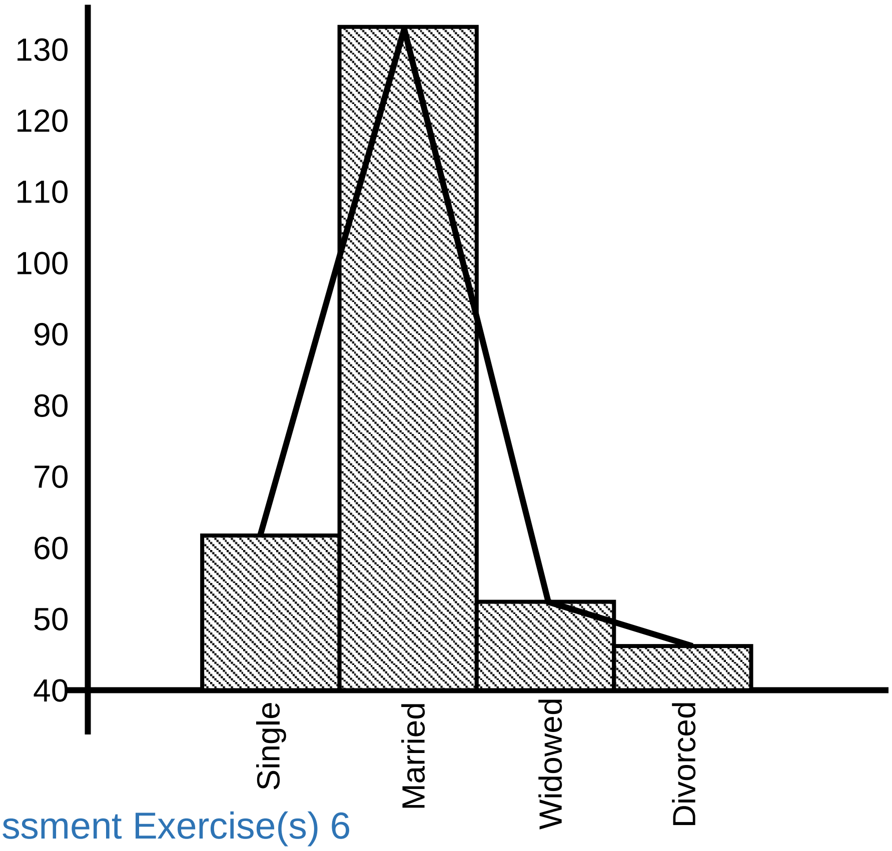
Histograms and bar charts look alike in presentation but while the bars of the bar charts are usually not joined; those of the histogram are usually joined. Further, while the chart attaches importance only to its heights, histogram attaches importance to both heights and the widths.

Self-Assessment Exercise 5.

1. Obtain the histogram of the data in the SAE 2 above.

3.2.4 Frequency polygon

A frequency polygon is obtained by joining the mid-points of the tops of the rectangles of a histogram.
Example: Draw the frequency polygon for the example above.



Self-Assessment Exercise(s) 6

1. Of 100 patients in an orthopaedic hospital who were asked for their room, 50 wanted private rooms, 40 wanted semi-private and 10 would make do with any room. Present this data by means of a bar chart.

2. Of 400 nursing students at teaching hospital,152 planned too into psychiatric speciality,120 into paediatric, 80 into public health and 48 into orthopaedic nursing. Represent the information on pie chart.

3 In a study of age distribution of patients in orthopaedic hospital, the following ages were recorded:

51 35 45 52 53 32 31 44 47 35 52 36 44 45 44 32
48 44 44 33 53 44 44 47 44 44 44 55 44 34 54 44
45 48 32 44 47 58 50 37 44 47 50 46 38 57 49 50
51 38

Draw the frequency polygon for data above

4.0 Conclusion

In this unit, you have studied several graphical representations of data. These representations are used in interpreting features of data. They are descriptive in nature.

5.0 Summary

You have learned the following concepts in this unit.

1. The raw data resulting from a survey or census are usually unorganised
2. A collection of data must be organised and summarised so as to reveal the significant features.
3. A collection of data may be described by frequency tables, pie chart, bar chart, histograms and frequency polygons.

6.0 Tutor-Marked Assignment

1. Tabulation thus enhances the condensation and easy comparison of data (Yes or No)?
2. Of 100 patients in an orthopaedic hospital who were asked for their room,50 wanted private rooms, 40 wanted semi-private and 10 would make do with any room. Present this data by means of a bar chart.
3. Of 400 nursing students at teaching hospital, 152 planned too into psychiatric speciality,120 into paediatric, 80 into public health and 48 into orthopaedic nursing. Represent the information on pie chart.
4. In a study of age distribution of patients in orthopaedic hospital, the following ages were recorded:

51	35	45	52	53	32	31	44	47	35	52	36	44	45	44	32	48
44	44	33	53	44	44	47	44	44	44	55	44	34	54	44	45	48
32	44	47	58	50	37	44	47	50	46	38	57	49	50	32	51	38

Draw the frequency polygon for the data above.

7.0 References/Further Reading

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

Indira Gandhi National Open University, (1999) Probability and Statistics, Sita Fine Arts Pvt. Ltd., New Delhi-28

Module 3

Measures of Central Tendency

Unit 1: Measures of Location

Unit 2: Weighted Mean

Unit 1

Measures of Location

Content

1.0 Introduction

In population as well as in sample data sets, there is the tendency for most of the observations to lie centrally within the given set of data arranged according to magnitude. An index to describe this concentration of values near the middle is customarily referred to as a measure of central tendency, a "typical" value, or simply, an average. It is called a measure of location because it indicates where, among the possible values of a variable, the population or sample is located. Measures of location are very useful parameters in that they describe a property of populations. Rather than compare entire distributions of sets of data with each other, it is usually more efficient to compare only certain characteristics of their parameters. We will discuss here, characteristics of some of these parameters and their associated sample statistics.

2.0 Learning Outcomes

Upon completion of this unit, you should be able to:

1. Apply use summation operator in computation
2. Define and compute the following measures of central tendency.
   a. Mean
   b. Median
   c. Mode.
3. Summarise data by means of central tendency.

3.0 Learning Content

3.1 Grouping and Tabulation

Sometimes the figures in a data are so spread that unless the figures are grouped, a neat and sensible frequency table may not be achieved. Tabulation done in this way is called a grouped frequency distribution table. The figures are usually grouped into distinct classes to avoid confusion of possible placement of data into two or more classes. So, classes may not have gaps in reality.

Self-Assessment Exercise 1

The weights in Kg of a collection of 40 workers in an organization are given below: 59, 53, 66, 55, 57, 65, 48, 59, 51, 58, 52, 68, 60, 70, 71, 56, 70, 64, 54, 67, 62, 53, 49, 56, 63, 48, 57, 61, 58, 55, 56, 55, 61, 52, 54, 65, 56, 50, 62, 60.

Using the tally method, prepare a grouped frequency distribution table using groups 48 – 52, 53 – 57, …

3.2 Mean (Averages)

Consider the statement:
"any average student should pass JAMB examination".

The word average' is used here to denote the not-too- brilliant and not-too-dull student. But in statistics the word has a special meaning. In the above context it would be used statistically, to describe that student who is representative, in some ways, of all students that sat for the examination. Therefore, if we have a group of figures, the average figure is that single figure that can represent all the other groups in that distribution. Three types of averages are often used in statistics:

i. The mean
ii. The median, and
iii. The mode

3.2.1 The Mean

For n numbers x_1, x_2... x_n, the mean, denoted by

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\bar{x} = \Sigma x / n$$

e.g. for the set 3, 5, 7, 10, 15.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{3 + 5 + 7 + 10 + 15}{5}$$
$$= 40/5$$
$$= 8$$

Note: When each of the numbers x_1, x_2,…,x_n has attached frequencies f_1, f_2,...,f_3, then the mean becomes

$$\text{Mean} = \bar{x} = \frac{\Sigma x}{\Sigma f}$$

Where $\Sigma f = n$

Self-Assessment Exercise(s) 2

1. The figures are usually grouped into distinct classes to avoid confusion of possible placement of data into two or more classes (Yes or No)?
2. A large and ungrouped data are cumbersome to study and interpret (Yes or No)?

3.2.2 Mean of a Grouped Data

Three methods of calculation are:

i. The long method
ii. The assumed mean method
iii. The coding method.

Long Method

$$\bar{x} = \varepsilon f x$$

$$= \frac{\Sigma f x}{\Sigma f}$$

Assumed Mean Method

$$x^{-} = A + \frac{\Sigma f d}{\Sigma f}$$

Where,

A is a guessed or assumed mean and d = X– A are the deviations from the assumed mean.

Coding Method

$$\bar{x} = A + \left(\frac{\Sigma f u}{\Sigma f}\right) + C$$

Where: A is an appropriate chosen x value,

C is the common class size and

u = …, -3, -2, -1, 0, 1, 2, 3…..

NOTE: The coding method is very short and should always be used for grouped data when class intervals are equal.

Self-Assessment Exercise 3

1. Referring to the table in the above SAE 1, calculate the mean using:
   a. The long method
   b. Assumed mean of 61
   c. The coding method.

Advantages of Mean

It takes account of all the values of a distribution. It is therefore, more representative than the other two and for this reason alone, it is used more than the other two averages.

Disadvantages of Mean

1. It is often the most difficult to calculate.
2. It is not easily understood by non-statistician.
3. While the mice and median often represent actual scores belonging to some members of the population, the. Mean often does not.
4. When the mean is used with discrete variables (e.g. number of children), it often yields unrealistic values such as 2.5 children.
5. While the median and mode can be obtained graphically the mean cannot.

Self-Assessment Exercise(s) 4

1. What are the merit and demerit of mean as a measure of location?
2. The distribution of the number of overtime hours per month worked by 60 staff of NITEL are given below:

Overtime (hrs)	60 – 69	50 – 59	40 – 49	30 – 39	20 – 29	10 – 19
No. of workers	5	11	17	14	9	4

a. Calculate the mean overtime hour using:

i. The long method
ii. Assumed mean of 40.5
iii. The coding method

3.3 The Median

If a set of data arranged in order of magnitude, the middle value, which divides the set into two equal groups is the median. Generally, for N data,

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

Example:

Find the median of the following sets of data
(a) 3, 6, 2, 4, 3
(b) 2, 5, 3, 4, 8, 3

Solution:

(a) Arrangement in order: 2, 3, 3, 4, 6

Here N= 5 and

$$\text{Median} = \left(\frac{5+1}{2}\right)^{th} \text{item}$$

$$= 6/2 = 3$$

That implies the 3rd item =3

(b) Arrangement in order: 2, 3, 3, 4, 5, 3

Here, N = 6

Thus, Median = $\left(\frac{6+1}{2}\right)^{th}$ item

= 3.5th item

This will be in interpreted as the $\frac{3^{rd}\text{ item} + 4^{th}\text{ item}}{2}$

Thus, median

3.3.1 Median of a grouped data

The median of a group data can be obtained graphically from the cumulative frequency curve (ogive) or by calculation, using the formula:

$$\text{Median} = L + \left(\frac{N/2 - F}{f}\right) C$$

Where:

L = Value of the lower-class boundary of the median class.
F = Cumulative frequency of the class just above the one containing the median.
f = Frequency of the median class
C = Size of the median class interval
NOTE: Usually we first obtain the value of N/2 which will enable us locate the position of the median in the cumulative frequency distribution.

Advantages of Median

1. It is easily understood.
2. It is relatively easy to calculate.

Disadvantages of Median

1. It takes no account of extreme values in the distribution. For instance, the median of 2, 40, 43, 45, and 96 is even though there are two extreme values 2 and 96.
2. It does not use all the data available.

Self-Assessment Exercise(s) 5

1. What are merit and demerit of median as a measure of central location?
2. The distribution of the number of overtime hours per month worked by 60 staff of NITEL are given below:

Overtime (hrs)	60 – 69	50 – 59	40 – 49	30 – 39	20 – 29	10 – 19
No. of workers	5	11	17	14	9	4

a. Construct the cumulative frequency curve and from it estimate the median.
b. Calculate the median and compare your results.

3.4 The Mode

This is the value or number that has the highest frequency in a distribution. The mode may not exist and even when it does exist, it may not be unique.

Example:

(a) 5, 2, 4, 7, 5, 3 has mode = 5 (unimodal)
(b) 2, 6, 3, 4, 3, 2, 5 has two modes 2 and 3 (bimodal)
(c) 4, 7, 2, 1, 3 has no mode.

The mode can be obtained both graphically and by calculations. For a grouped data, we use the histogram to estimate the mode while by calculation we use the formula:

$$\text{Mode} = \left(\frac{L + f_m - f_a}{2f_m - f_a - f_b} \right) C$$

Where:
L = lower class boundary of the modal class
f_m = frequency of the modal class
f_a = frequency of the class above the modal class
f_b = frequency of the class below the modal class
C = size of the modal class interval.

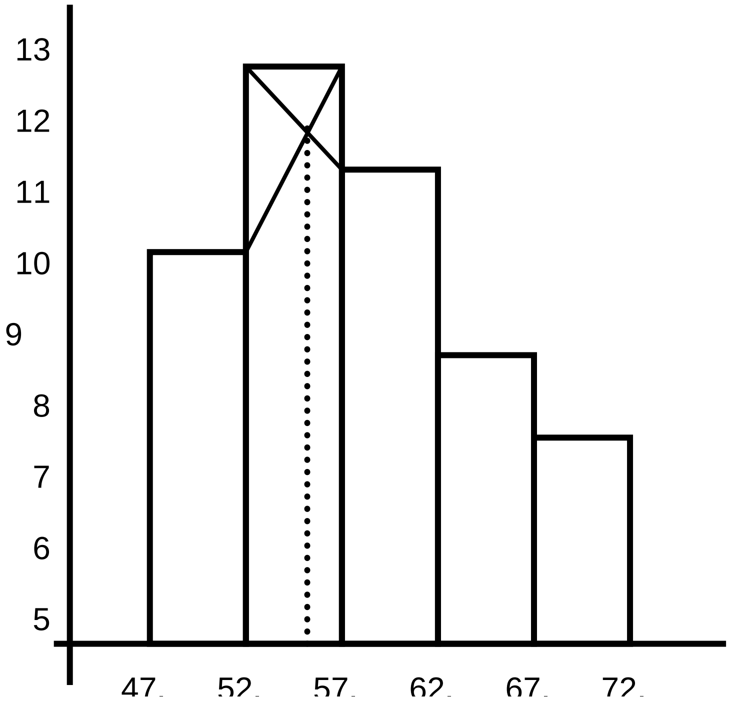
Note: The modal class is the class that has the highest frequency. The mode itself is a number within this class.

Referring to the table above

a. Construct the histogram and from it, estimate the mode of the distribution.
b. Calculate the mode and compare your answer with the estimated value in (a) above.

Solution:

a. The construction should be done on a graph sheet where frequencies are on the vertical axis and class boundaries on the horizontal



The mode is approximately 56.

b. Mode = $L + \left[\frac{f_m - f_a}{2\,f_m - f_a - f_b}\right] c$

Here, the modal class is 53 – 57.

Hence, L = 52.5, fm = 12, fa = 10 and C = 5

Mode = $52.2 + \left[\frac{12-8}{2(12) - 8 - 10}\right] \text{x } 5$

= 52.5 + (4/6) x 5

= 52.5 + 3.33

= 55.83

Comparison: Graphical value = 56

Estimated value = 55.83

The values agree appreciably.

Advantages of Mode

It is often the easiest to calculate of the three.

Disadvantages of Mode

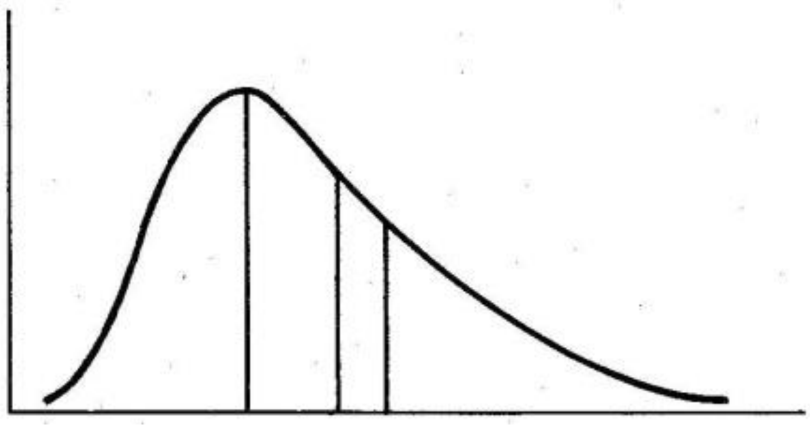
1. It presents a misleading picture for a distribution that does not have a regular shape.
2. It does not use all the data available.

3.4.1 Relation between The Mode, Median and Mean

For unimodal frequency curves which are moderately skewed, the following relation between the mean, median and mode holds:

$$\boxed{\textbf{Mean} - \textbf{Mode} = 3(\textbf{Mean} - \textbf{Median})}$$

The figure below shows the relative positions of the mean, median and mode for frequency curves which are skewed to the right and left respectively. For symmetrical curves the mean, mode and median all coincide.



Self-Assessment Exercise 6

1. The distribution of the number of overtime hours per month worked by 60 staff of NITEL are given below:

Overtime (hrs)	60 – 69	50 – 59	40 – 49	30 – 39	20 – 29	10 – 19
No. of workers	5	11	17	14	9	4

a. Calculate the mode of the distribution.

4.0 Conclusion

In this unit, you have been exposed to the measure of central tendency which are bench marks, typical scores or measures which give precise and brief description of a set of data. These are very important aspect of statistics you cannot toy with.

To make your data very precise for interpretation, you will need to learn these measures of location very well.

5.0 Summary

In this unit you have learnt that the measures of central tendency are a set of bench marks which make precise and brief presentation or description of a set of scores. The three basic measures of central tendency are the mean, the median and the mode.

The mean is the most widely used. It is equal to the sum of the scores divided by the number of the scores. The symbol is $\bar{x}$ and the formula is ${}^{\Sigma X}/_{N}$ or ${}^{\Sigma FX}/_{\Sigma F}$. Or for assumed mean = AM + int(Σfx/Σfx).

6.0 Tutor Marked Assignment (TMA)

Use the data below to find;

a. Mean
b. Median and
c. Mode

S/N	CLASS INTERVAL	F
1	75-79	2
2	70-74	4
3	65-69	6
4	60-64	10
5	55-59	25
6	50-54	35

7	45-49	20
8	40-44	15
9	35-39	10
10	30-34	5

7.0 References/Further Reading

Hannagan, T.J. (1982) Mastering Statistics, New York: The Macmillan Press Ltd
Ary Donald and Jacobs, L.C. (1996) Introduction to Statistics,

Unit 2

Weighted Mean

Content

1.0 Introduction

Consider this situation, students are admitted to a B.Sc. course in statistics on the basis of their performances in the Higher Secondary, or an equivalent examination. Should the scores in mathematics papers should be considered more important than those in physics papers? Similarly, should the scores in language papers not be least important? It is necessary in such a situation to take into account the relative importance (or weight) of the different observations while evaluating the mean.

2.0 Learning Outcomes

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

1. Calculate the weighted mean (arithmetic, geometric and harmonic means)
2. Discuss the relationship between arithmetic, geometric and harmonic Means

3.0 Learning Content

3.1 Weighted Arithmetic Mean

Sometimes we associate with the numbers X_1, X_2, . . ., X_n certain *weighting factors* or *weights* w_1, w_2, . . ., W_n depending on the significance or importance attached to the numbers. In this case

$$\bar{x} = \frac{w_1X_1 + w_2X_2 + \ldots + W_nX_n}{W_1+W_2+\ldots+W_K} \quad = \sum wX$$

is called the *weighted arithmetic mean.*

Example: 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 grades 70 and 90, the mean grade is:

$$\bar{x} = \frac{(1)(70) + (1)(90) + (3)(85)}{1+1+3}$$

$$= 415/5$$

$$= 83$$

Properties of the arithmetic mean

1. The algebraic sum of the deviations of a set of numbers from their arithmetic mean is zero.

Example: The deviations of the numbers 8,3, 5, 12, 10 from their arithmetic mean 7.6 are 8-7.6, 3-7.6, 5-7.6, 12-7.6, 10-7.6

= 0.4, -4.6, -2.6, **4.4**, 2.4 with algebraic sum
0.4-4. 6-2.6 + 4.4 + 2.4 = 0.

2. The sum of the squares of the deviations of a set of numbers, X, from any number *a* is a minimum if and only if *a = X.*

   i) e.g. Prove that $W^2 + pw + q$, where p and q are given constants, is a minimum if and only if W = - ½P

      using (a), show that $\frac{\sum(x - a)^2}{N}$

   ii) Prove that $\frac{\sum(x - a)^2}{N}$ is a minimum if and only if $a = \bar{x}$

Solution:

$$\frac{\sum(x^2 - 2ax + a^2)}{N} = \frac{\sum x^2 - 2a\sum x + Na^2}{N}$$

$$= a^2 - \frac{2a\sum x}{N} + \frac{\sum x^2}{N}$$

Comparing the last expression with $(w^2 + pw + q)$, we have:

W=a, p= -2 ∑x/N, Q = $\sum X^2/N$

Then the expression is a minimum when a=-1/2p = ∑x/N = X

(c) If f_1 numbers have mean m_1, f_2 numbers have mean m_2, . . ., f_K numbers have mean m_K then the mean of all the numbers is

$$= \frac{f_I m_i + f_2 m_2 + \ldots + fkmk}{f_I + f_2 + \ldots + f_k}$$

i.e. a weighted arithmetic means of all the numbers

E.g. if a company having 80 employers, 60 earn ₦3 per hour and 20 earns ₦2 per hour, determine:

(i) The mean earnings per hour

(ii) Would the answer to (a) be the same if the 60 employers earn a mean hourly wage of ₦3 per hour and the 20 employers earn a mean hourly wage of ₦2 per hour?

(iii) Do you believe the mean hourly wage to be typical?

Solution

(i) $\bar{x} = \sum fx/\sum f$

$$= \frac{60x3 + 20 \times 2}{60 + 20}$$

= 220/80

= ₦2.75

(ii) Yes, the result is the same

(iii) Yes, it is typical.

(d) If A is any guessed or assumed arithmetic mean (which may be any number) and if = X-A, denoted by d, are the deviations of X from A, then,

$$x = A + \sum di/N$$

Or simply $X = A + \sum d/N$

If the data is grouped, then

$\bar{x} = A + \sum f_i d_i / \sum f_i$ or simply as

$\bar{x} = A + \sum fd/N$
where $N = \sum f_i$

Self-Assessment Exercise(s) 1

1. The algebraic sum of the deviations of a set of numbers from their arithmetic mean is zero (Yes or No)?
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 grades of 70 and 90, calculate the mean grade.

3.2 Geometric Mean and Harmonic Mean

3.2.1 Geometric Mean

The geometric mean, G of a set of N numbers $X_1, X_2, X_3, ., X_N$ is the Nth root of the product of the numbers:

$$G = N\ X1.X2.X3\ldots X_N$$

E.g. The geometric mean of the numbers: 2, 4, 8 is:

$$3\ (2)\ (4)\ (8) = \sqrt[3]{64}$$
$$=8$$

In practice, G is computed by logarithms. For the geometric mean from grouped data, use the mid- class intervals as your values.

3.2.2 The Harmonic Mean, H

The harmonic mean, *H* of a set of *N* numbers: $X_1, X_2, X_3, \ldots, X_N$, is the reciprocal of the arithmetic mean of the reciprocals of the numbers:

$$H = N / \sum 1/x$$

Example: The harmonic means of the numbers *2,* 4, 8 is

$$H = \frac{3}{½ + ¼ + 1/8}$$

$$= \frac{3}{7/8}$$

$$= 3.43$$

NOTE: It is often convenient to express the fractions in decimal form first.

Self-Assessment Exercise(s) 2

1. The harmonic mean, *H* of a set of *N* numbers: X_1, X_2, X_3, .. . , X_N , is the reciprocal of the arithmetic mean of the reciprocals of the numbers (Yes or No)?
2. Calculate the geometric mean of the numbers: 3, 5, 6, 7.
3. Calculate harmonic mean of the numbers: 5, 6, 7.

3.3 Relation between Arithmetic, Geometric, and Harmonic Means

The geometric mean of a set of positive numbers X_1, X_2, ..., X_r is less than or equal to their arithmetic mean but is greater than or equal to their harmonic mean. In symbols,

$$H \leq G \leq X$$

The equality signs hold only if all the numbers X_1, X_2, . . . , X are identical.

Example: The set 2, 4, 8 has arithmetic mean 4.67, geometric mean 4, and harmonic mean 3.43.

Example: During one year the ratio of milk prices per quart to bread prices per loaf was 2. Whereas, during the next year the ratio was 2.00.

i. Find the arithmetic mean of the ratios for the two year period.
ii. Find the arithmetic mean of the ratios of bread prices to milk prices for the two year period.
iii. Discuss the advisability of using arithmetic mean for averaging ratios.
iv. iv. Discuss the suitability of the geometric mean for averaging ratios.

Solution:

i. Mean ratio of milk to bread prices = 1(2.50+2.00) 2.25

ii. Since the ratio of milk to bread prices for the first year is 2.50, the ratio of bread to milk price is 1/2.50 = 0.40. Similarly, the ratio of bread to milk prices for the second year is 1/2.00 0

Then

Mean ratio of bread to milk prices = 1(0.40+0.50) = 0.45

iii. We would expect the mean ratio of milk to bread prices to be the reciprocal of the mean ratio bread to milk prices if the mean is an appropriate average.

However, 1/0.45 = 2.11 # 2.25.
This shows that the arithmetic mean is a poor average to use for ratios.
iv. Geometric mean of ratios of milk to bread prices $= \frac{\underline{V(2.50)(2.00)}}{V\ 5}$

Geometric mean of ratios of bread to milk prices v'(0.40)(0.50) = V 0.2 = 1/ V 5 Since these averages are reciprocals, our conclusion is that the geometric mean is more suitable than the arithmetic mean for averaging ratios for this type of problem.

Self-Assessment Exercise 3

1. Explain the relation among arithmetic, geometric and harmonic means.

4.0 Conclusion

We have been exposed to the concept of weighted mean: be it arithmetic, geometric and harmonic means. This means for us to calculate appropriate mean, weight has to be attached to individual value accordingly.

5.0 Summary

In this unit, you have been learnt that the concept of weighted mean is particularly useful in the construction of price index number and in such a situation to take into account the relative importance (or weight) of the different observations while evaluating the mean.

6.0 Tutor Marked Assignment (TMAs)

1. The algebraic sum of the deviations of a set of numbers from their arithmetic mean is zero (Yes or No)?
2. The price increases from 2010 to 2011for five food items have been (in percentage terms) as follows:

   132.1 153.4 144.3 119.7 120.1

   And the relative importance of these items in a typical citizen's diet is:

   34 19 24 12 11

   Calculate the average price increase for these items.
3. The harmonic mean, *H* of a set of *N* numbers: X_1, X_2, X_3,. , X_N, is the reciprocal of the arithmetic mean of the reciprocals of the numbers (Yes or No)?
4. Calculate the geometric mean of the numbers: 2, 4, 6, 8.

5. Calculate harmonic mean of the numbers: 2, 4, 6, 8.
6. Explain the relation among arithmetic, geometric and harmonic means.

7.0 References/Further Reading

Harper W.M. (1982) Statistics, Fourth Edition. Macdonald and Evans Handbook Series.

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

Module 4

Measures of Dispersion

Unit 1

Measures of Dispersion 1

Content

1.0 Introduction

The degree to which numerical data tends to spread about an average value is called the variation or dispersion of the data. Consider the following distribution of wages in Naira of 5 workers in each of two television stations A and B.

Station A	25,000	30,000	35,000	40,000	45,000
Station B	7,500	12,500	35,000	55,000	65,000

The mean and the median wages for each of the two distributions is N35, 000. From the results, one could wrongly conclude that the workers conditions of service in both stations are the same. A close observation of the figures clearly shows that the wages of workers in station A are more fairly and evenly distributed than those in B. One therefore, needs a study of dispersion to detect the disparity in a distribution. Various measures of dispersion are available; the measures which we shall discuss in this unit are the range, the quartile, the deciles and percentiles.

2.0 Learning Outcomes

By the end of this unit you will be able to:

1. Define and calculate the range in a given set of scores.
2. Explain and locate the quartiles in a distribution of scores.
3. Explain and locate the deciles in a set of scores
4. Explain and calculate the percentiles in a given set of scores.

3.0 Learning Content

3.1 The Range

This is the simplest but crude and unreliable method of estimating variability. It is defined as the difference between the highest and the lowest scores in a given distribution. It is usually affected by the presence of two extreme scores. The greater the range, the greater is the dispersion or variability. It can be found by using the formula:

$R = X_h - X_l$,

where X_h represents the highest score and X_l is the lowest scores.

Example:

Find the range in the following set of scores.

53, 59, 72, 62, 57, 54, 66, 79, 14, 65, 64, 95, 59.

If you look at the scores very well, you will notice that the lowest score $X_l = 14$ and the highest score $X_h = 95$. Therefore the range R. will be :

$X_h - X_L = 95-14$

$= 81.$

Self-Assessment Exercise 1

1. Find the range in the set of scores below

   53, 59, 60, 48, 64, 72, 56, 34, 75, 52, 36, 93

3.2 Quartiles

In the last unit, you learnt that Median is a positional score, which occupy the middle point on the score scale. In the same way, the quartiles are positional scores. The first Quartile Q_1 is the score point that sets the lower quarter or 25% of the group. In the same way, the middle quartile Q_2 is the median score point and third quartile Q_3 is the 75% of the group. So, quartiles are points that divide a score into four equal parts. These points can locate in a distribution.

Quartile deviation or the semi-interquartile range

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

Where Q_1 and Q_3 are the lower and upper quartiles respectively. Example: Locate the Q_1 and Q_3 in the data given below.

Scores	15	18	21	23	25	27	28	29	32
Freq.	1	1	2	3	6	5	3	4	3

Solution:

i. Complete the table to include the cumulative frequency.

S/No	Scores	F	CF
1	32	3	28
2	29	4	25
3	28	3	21
4	27	5	18
5	25	6	13
6	23	3	7
7	21	2	4
8	18	1	2
9	15	1	1

ii. Find the 25% or ¼ of the number of scores $={}^{25}/_{100} \times {}^{28}/_{1}=7$

iii. Count below along the frequency column until you get 25% of the cases. It is between 23 and 25 i.e Q_1= $^{23+25}/2$ =24

iv. Find 75% or ¾ of the scores =$^{75}/_{100}$x$^{28}/_{1}$ =21

v. Count from the below along the frequency until you get 75% of the cases. This gives Q_3. It is between 28 and 29 i.e Q_3=$^{28+29}/2$=28.5

Self-Assessment Exercise 2

1. The table below shows a frequency distribution of weekly wages in naira of 65 employees at the P and R Company, locate the Q_1 and Q_3 in the data given below.

Wages (₦)	50-59.99	60-69.99	70-79.99	80-89.99	90-99.99	100-109.99	110-119.99
No of Employees	8	10	16	14	10	5	2

3.3 The Deciles and the Percentiles

3.3.1 The Deciles

In this sub unit, we shall move to another step. This is to divide into ten equal parts to locate the deciles. Deciles points are used to mark off a distribution, thus indicating points of dividing a distribution of success into tenths. Thus, there are 9 deciles i.e. from 1 to 9 which divide a distribution into ten equal parts. D_1 is the first deciles and below D_1 lies the bottom 10% of the group. In the same way D_2 is the point in the distribution below which 20% of the cases fall. Like quartiles, deciles are points in a distribution not segments.

3.3.2 The Percentiles

Percentiles are ordinal measures. They are scores points which divide the distribution into 100 equal parts called percentages. In other words, they are points on the raw scale below which given percentages of the cases in the distribution fall. For instances, 80th percentile is the point on the score scale that has exactly 80% of the cases below it.

Percentiles are symbolised by the letter P_x, with x denoting the particular percentile. Thus 90th percentile is written P_{90}, they are used for decision making when part of a population is to be selected because of its position within the total.

Note that the median corresponds to the 50th percentiles, P_{50} and 2nd Quartile Q_2

The 1st quartile corresponds to the 25th percentile, P_{25}

The 3rd quartile corresponds to the 75th percentile, $_{P75}$

Self-Assessment Exercise(s) 3

1. Locate the following in the data given in the Self-Assessment Exercise 2 above:

a. D_2

b .D_4

2. Explain what do you understand by percentile.

Scores	16	18	20	22	24	26	28	30	32
Freq.	1	1	2	3	6	5	3	4	3

4.0 Conclusion

In this unit, we have gone through sources of the measures of variability or dispersion. These are the measures used to establish the homogeneity or heterogeneity of a set of data in a distribution scale.

5.0 Summary

In this unit you have been exposed to some measures of variability which are measures that show the spread of the scores in a given distribution. The measures you have seen so far are:

i. The range which simply shows the difference between the highest and the lowest observations or numbers.

ii. The quartiles are the points which divide the distributions or scores into four equal parts called quarters.

iii. The deciles are also points on the distribution that divide the distribution into ten equal parts or tenths.

iv. Percentile are points on the score scale that divide the distribution into 100 equal parts called centiles or percentages.

6.0 Tutor-Marked Assignment

1. Find the range in the set of scores below

   53, 59, 60, 48, 64, 72, 56, 34, 75, 52, 36, 93

2. i. Locate the Q_1 and Q_3 in the data given below.

   ii. Find semi interquartile range.

3. Explain the following:

i. P_{50}

ii. P_{95}

iii. D_2

iv. D_4

7.0 References/Further Reading

Harper W.M. (1982) Statistics, Fourth Edition. Macdonald and Evans Handbook Series.

Ary, Donald and Jacobs (1976) Introduction to Statistics: Purposes and Procedures, New York.

Unit 2

Measures of Dispersion 2

Content

1.0 Introduction

In unit 4, you learnt the various measures of central tendency and you are told that these measures are a set of bench marks which make precise and brief presentation or description of a set of that. Therefore, in order to describe a distribution adequately, we shall need both the measures of central tendency and variability. This is because information concerning variability may be as important as or more important than information concerning central tendency.

2.0 Learning Outcomes

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

1. Calculate the mean deviation
2. Calculate the variance in a set of scores.
3. Calculate the standard deviation in a given distribution

3.0 Learning Content

3.1 Mean Deviation

At the aforementioned measures of dispersion only use two
points in the distribution; they are therefore statistically unreliable. A measure which makes use of all the available data is the mean deviation.

$$Mean\ deviation\ = \frac{\sum|x - \bar{x}|\ for\ discrete\ (ungroup)\ data.}{n}$$

However, when frequencies are attached, then

$$\text{Mean deviation} = \frac{\Sigma\ f|x - \bar{x}|}{\Sigma f}$$

Where, n = Σ f and X = the mean.

Example: Find the mean deviation of the set of values: 2, 4, 7, 10, 12

Solution

$$\text{Mean deviation} = \frac{\Sigma\ f|x - \bar{x}|}{\Sigma f}$$

$$\text{Here, } X = \frac{\Sigma\ f}{n} = \frac{2 + 4 + 7 + 10 + 12}{5}$$

= 35/5

= 7

Thus:

X	**$x - \bar{x} = x - 7$**	**$\lvert x - \bar{x}\rvert = \lvert x - 7\rvert$**
2	-5	**5**
4	-3	**3**
7	0	**0**
10	3	**3**
12	5	**5**
Total		**16**

Hence:

Mean Deviation = 16/5 = 3.2

Note: For a grouped data, the class marks are taken as our x values.

Self-Assessment Exercise 1

1. Find the mean deviation of the set of values: 12, 6, 7, 3, 15, 10, 18, 5

3.2 Variance and Standard Deviation

This is the most reliable, useful and mostly used measure of dispersion. Reasons for this are not farfetched.

1. The standard deviation makes use of all the members of a distribution.

2. It yields itself for further statistical usage as used in computations under normal distribution.

Calculations

1. Long Method:

$$S = C \times \sqrt{\frac{\Sigma fx^2}{\Sigma f} - (\frac{\Sigma fx}{\Sigma f})^2}$$

2. Assumed Mean Method:

$$S = C \times \sqrt{\frac{\Sigma fd^2}{\Sigma f} - (\frac{\Sigma fd}{\Sigma f})^2}$$

Where d = x – A

3. Coding Method:

$$S = C \times \sqrt{\frac{\Sigma fu^2}{\Sigma f} - (\frac{\Sigma fu}{\Sigma f})^2}$$

Where **u** is as earlier defined under mean.

Note: In problems involving calculation of both mean and standard deviation, for simplicity, the method used in computing the mean should be applied to find the standard deviation.

The square of the standard deviation is called the variance. So if 'S' denote the standard deviation, then S^2 is the Variance.

Examples:

1. The marks scored by some 50 students in a statistics test are given below:

Marks	51 – 60	41 – 50	31 - 40	21 - 30	11 – 20	1 – 10
Frequency	5	10	11	12	6	6

a. Calculate the mean and the standard deviation using the log method

Solution:

	F	**X**	**Fx**	**x^2**	**fx^2**	**$x-\bar{x}$ = x-31.1**	**|x-31.1|**	**f|x-31.1|**	**F**
1 – 10	6	5.5	33.0	30.25	181.5	-25.5	25.5	153.6	6
11 – 20	8	15.5	93.0	240.25	1441.5	-15.5	15.5	93.5	12
21 – 30	12	25.5	306.0	650.25	7803.0	-5.5	5.5	67.2	24
31 – 40	11	35.5	390.5	1260.25	13862.75	4.4	4.4	43.4	35
41 – 50	10	45.5	455.0	2070.25	20702.25	14.4	14.4	144.0	45
51 – 60	5	55.5	277.5	3080.25	14401.25	24.4	24.4	122.0	50
Total	**50**		**1555.0**		**59392.5**			**628.8**	

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{1555}{50}$$

$$= 31.1$$

a. $$S = C \times \sqrt{\frac{\Sigma fu^2}{\Sigma f} - (\frac{\Sigma fu}{\Sigma f})^2}$$

$$S = \sqrt{\frac{59392.5}{50} - (\frac{1555}{50})^2}$$

$$= 14.85$$

Self-Assessment Exercise 2

1. The distribution of the ages of 100 people in a village is given below:

Ages	60- 62	63-65	66 -68	69-71	72- 74
Frequency	5	18	42	27	8

Using assumed mean 67, calculate the standard deviation and the variance of the distribution.

4.0 Conclusion

In this unit you have learnt that apart from the usefulness of the measures of the measures of central tendency for providing a concise index of the average value of set scores, three is more to be studied about a set of scores. Therefore, to describe a distribution of scores very well and adequately we need both the measures of central tendency and measure of variability, this is because the two measures make up two types of descriptive statistics which are indispensable in describing distribution of a given data.

5.0 Summary

You have studied three measures in this unit; they are mean deviation, standard deviation and variance. Thus,

Mean Deviation = $\frac{\Sigma f|x - \bar{x}|}{\Sigma f}$

Standard Deviation

i. Long Method:

$$S = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - (\frac{\Sigma fx}{\Sigma f})^2}$$

ii. Assumed Mean Method:

$$S = \sqrt{\frac{\Sigma fd^2}{\Sigma f} - (\frac{\Sigma fd}{\Sigma f})^2}$$

Where d = x – A

iii. Coding Method:

$$S = C \times \sqrt{\frac{\Sigma fu^2}{\Sigma f} - (\frac{\Sigma fu}{\Sigma f})^2}$$

6.0 Tutor-Marked Assignment

1. Find the mean deviation of the set of values: 2, 3, 5, 6, 8.
2. The distribution of the ages of 108 staff of a telecommunication outfit are given below:

Ages	15-19	20-24	25-29	30-34	35-39	40-44	45-49
No of Staff	8	12	20	24	16	16	12

Using assumed mean 33, calculate the mean and the standard deviation of the distribution.

7.0 References/Further Reading

Murry R. Spiegel, Outline of Theory and Problems of Statistics (1961),Schaum Publishing Company. U.S.A.

Harper W.M (1982) Statistics, Fourth Edition, Macdonald and Evans.

Module 5

Skewness and Kurtosis

Unit 1

Skewness and Kurtosis

Content

1.0 Introduction

In unit 4, we talked about the central tendency and in the last two units we discussed measures of dispersion. Now, in the unit, we shall discuss two additional features of frequency distributions. These are: skewness and kurtosis. A measure of skewness would tell us how far the frequency curve of the given frequency distribution deviates from a symmetric one. On the other hand, a measure of kurtosis gives us some information about the degree of flatness (or peakness) of the frequency curve.

2.0 Learning Outcomes

At the end of this, you would be able to:

1. Compute some measures of skewness and kurtosis
2. Discuss the relative advantages and disadvantage of these measures

3.0 Learning Content

3.1 Skewness

Frequency distribution may be classified as symmetrical and skewed (or asymmetrical). Skewed distribution can also be classified as positively skewed or negatively skewed.

Now, the degree of skewness is the extent to which the given distribution departs from symmetry, a good measure of the degree of skewness has to fulfill the following criteria:

i. It should be a pure number; i.e. should be free of units in which the variable is measured.

ii. Should be zero, positive and negative for a symmetrical distribution, a positively skewed distribution and a negatively skewed distribution, respectively.

iii. It should vary between two definite limits, say, -k and +k, as the nature of a distribution changes from extreme negative asymmetry to extreme positive asymmetry.

Here are some commonly used measures (assuming $s > 0$). They include: the Pearson’s first and second co-efficient of skewness

i. $$\text{Skewness} = \frac{(\text{Mean- Mode})}{\text{Standard Deviation}}$$

ii. $$\text{Skewness} = \frac{3(\text{mean - median})}{\text{Standard Deviation}}$$

iii. $$\text{Quartiles coefficient of skewness} = \frac{(Q_3-Q_2) - (Q_2-Q_1)}{Q_3-Q_1}$$

iv. 10-90 percentile of skewness = $\frac{(P_{90}-P_{50})-(P_{5o}-P_{10})}{(P_{90}-P_{1o})}$

Example

1. The marks scored by some 50 students in a statistics test are given below:

Marks	51 – 60	41 – 50	31 - 40	21 – 30	11 – 20	1 – 10
Frequency	5	10	11	12	6	6

i. Calculate the mean and the standard deviation using the long method.

ii. Calculate the median of the distribution.

iii. Using your results in (a) and (b) above, compute an appropriate measure of skewness and hence comment on the shape of the distribution.

	F	**X**	**Fx**	**x²**	**fx²**	**x-X** **= x-31.1**	**|x-31.1|**	**f|x-31.1|**	**F**
1 – 10	6	5.5	33.0	30.25	181.5	-25.5	25.5	153.6	**6**
11 – 20	8	15.5	93.0	240.25	1441.5	-15.5	15.5	93.5	**12**
21 – 30	12	25.5	306.0	650.25	7803.0	-5.5	5.5	67.2	**24**
31 – 40	11	35.5	390.5	1260.25	13862.75	4.4	4.4	43.4	**35**
41 – 50	10	45.5	455.0	2070.25	20702.25	14.4	14.4	144.0	**45**
51 – 60	5	55.5	277.5	3080.25	14401.25	24.4	24.4	122.0	**50**
Total	50		1555.0		59392.5			628.8	

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{1555}{50}$$

$$= 31.1$$

$$\text{i. } S = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - (\frac{\Sigma fx}{\Sigma f})^2}$$

$$S = C \times \sqrt{\frac{59392.5}{50} - (\frac{1555}{50})^2}$$

$$= 14.85$$

ii. Median = L + $\underline{N/2 - F}$ C f

N/2 = 50/2 = 25,

Median class is 31 – 40

Thus, L = 30.5, F = 24, f = 11, C = 10.

$$\text{Median} = 30.5 + \frac{25 - 24}{11}\ \ 10$$

$$= 30.5 + 0.91$$

$$= 31.41$$

iii. The results in (a) and (b) above are: Mean = 31.1.

Standard deviation = 14.85, Median = 31.41.

Hence, the appropriate measure of skewness is the Karl Pearson's Coefficient of Skewness, which is given by:

$$\text{KPCS} = \frac{3(\text{mean} - \text{median})}{\text{Standard deviation}}$$

$$= \frac{3(31.1 - 31.41)}{14.85}$$

$$= \frac{-0.93}{14.85}$$

$$= -0.06$$

Comment: The value -0.06 is very close to zero, so we can say the shape of the distribution is almost symmetrical. (or we say, it is very moderately skewed to the left).

Self-Assessment Exercise 1

1. The table below shows a frequency distribution of weekly wages in naira of 65 employees at the P and R Company;

Wages (₦)	50-59.99	60-69.99	70-79.99	80-89.99	90-99.99	100-109.99	110-119.99
No of Employees	8	10	16	14	10	5	2

a. Calculate the mean, median, mode and coefficient of skewness of the distribution.

b. Using your results in (a) above, comment on the shape of the distribution.

3.2 Kurtosis

We now focus our attention on another feature of a frequency distribution that determines the shape of the distribution. It is the degree of steepness or pointedness of distribution. Some distributions are flat –topped; some are highly peaked; most distributions will be in between these two extreme types, not too peaked and not too flat-topped either.

The quantitative indices of kurtosis of a distribution can be calculated using the semi-inter quartile range and the ninetieth and tenth percentiles. The index is symbolized by the Greek letter Ϗ (kappa) and is given by:

$$\text{Ϗ}= \frac{\frac{1}{2}(Q_3-Q_1)}{(P_{90}-P_{10})} \quad \text{or} \quad \frac{(P_{75}-P_{25})}{(P_{90}-P_{10})}$$

Self-Assessment Exercise 2

1. Explain the following:
   a. Skewness
   b. Kurtosis

4.0 Conclusion

In this unit, you have learnt how to find out the degree to which a set data is skewed, and to find out spread or bunched of set scores which is technically referred to as Kurtosis.

5.0 Summary

In this unit, we have been able to discover that frequency distribution may be classified as symmetrical and skewed (or asymmetrical). Also, Skewed distribution can also be classified as positively skewed or negatively skewed. Another feature of frequency distribution is the kurtosis which is the degree of steepness or pointedness of distribution.

6.0 Tutor-Marked Assignment

The table given below shows the frequency distribution of the ages in years of 65 employees of an establishment.

Ages	20 -24	25 – 29	30 – 34	35 – 39	40 – 44	45 - 49
Frequency	8	10	16	14	10	7

i. Calculate Q_1, Q_2, Q_3, P_{10} and P_{90}

ii. Use the values you have calculated in (a) above to find: Quartile coefficient of skewness.

10 – 90 percentile coefficients of skewness.

Comment on the shape of the distribution.

7.0 References/Further Reading

Murry r. Spiegel, outline of theory and problems of statistics (1961), schaum publishing company. U.S.A.

National open university of nigeria (2006), business statistics 1

Module 6

Simple Regression

Unit 1: Dependent and Independent Variables

Unit 2: Simple Regression

Unit 1

Dependent and Independent Variables

CONTENT

1.0 Introduction

There are many statistical investigations in which the main objective is to determine whether a relationship exists between two or more variables. If such a relationship can be expressed by a mathematical formula, we will then be able to use it for the purpose of making predictions. The reliability of any prediction will, of course, depend on the strength of the relationship between the variables included in the formula.

If to each value which a variable X can assume, there are corresponds one or more values of variable Y, we say that Y is a function of X and write Y=F(x) (read 'Y equals F of X').

2.0 Learning Outcomes

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

1. Explain and differentiate between dependent and independent variables.
2. Explain in words the use of regression analysis.
3. Draw a scatter diagram corresponding to the given bivariate frequency distribution.

3.0 Learning Content

3.1 Dependent and Independent Variables

If to each value which a variable X can assume, there are corresponds one or more values of variable Y, we say that Y is a function of X and write Y=F(x) (read 'Y equals F of X'). To indicate this functional dependence, the variable X is called the independent variable and Y is called the dependent variable. If only one value of Y corresponds to each value of X, we call Y a single-valued function of X; otherwise it is called multiple-valued function of X.

The functional dependence or corresponding between variables is often depicted in a table. However, it can also be indicated by an equation connecting the variables, such as Y=2X-3, from which Y can be determined correspondingly to various values of X.

If Y=F(X), it is customary to let F(3) denote 'the value of Y when X=3', F(10) denote 'the value of Y when X=10' and so on.

Thus, if $Y=F(x) =X^2$, then $F (3) =3^2=9$ is the value of Y when X =3.

Consider Z=16+4X-3Y. Given the values of X and Y, there is a correspond value of Z. We can denote this dependence of Z on X and Y by writing Z=F(X,Y), read 'Z is a function of X and Y'. F(2,5) denotes the value of Z when X=2 and Y=5 and is 9. Similarly, F(-3,-7)= 25 and F(-4, 2)= -6.

The variables X and Y are called independent variables and Z is called dependent variable.

Self-Assessment Exercise(s) 1

1. Explain the following:
   i. Dependent variable X
   ii. Independent variable

2. Find A if $A = F(\pi r^2)$ and π =22/7, r= 7.

3.2 What Do We Mean by Regression Analysis?

Scientists, economists, psychologists and sociologists have always been concerned with the problems of prediction. A mathematical equation that allows us to predict values of one or more independent variables is called a Regression Equation.

For example, suppose X and Y denote respectively, the height and weight of adult males. Then a sample of N individuals would reveal the heights X_1,X_2,…,X_N and the corresponding weights Y_1, Y_2,...,Y_N. A next step is to plot the points (x_1,y_1), (x_2, y_2), …, (x_N, y_N) on a rectangular co-ordinate system. The resulting set of points is sometimes called a scatter diagram.

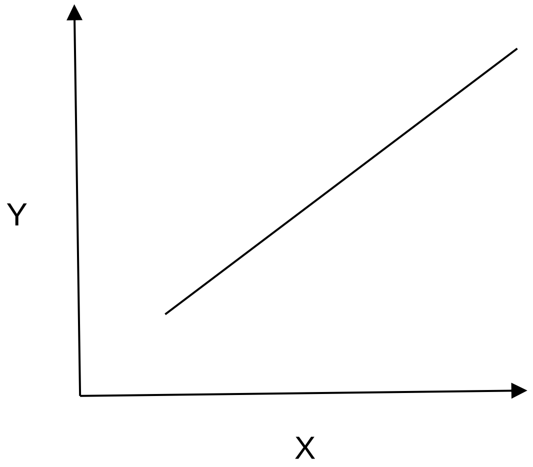
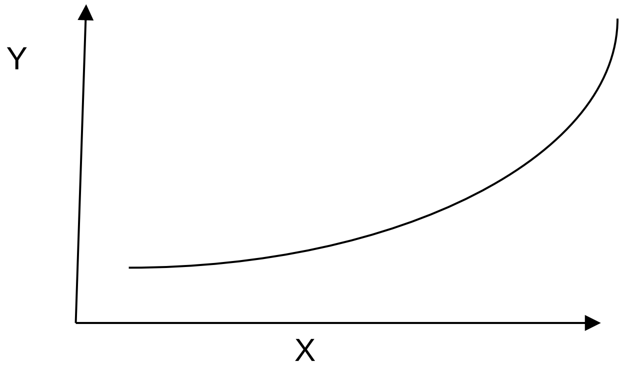


Fig (a)



From the scatter diagram, it is often possible to visualize a smooth curve approximating the data. Such a curve is called an approximating curve. In fig (a), the data appear to be approximated well by a straight line and we say that a linear relationship exists between the variables. In fig (b), however, although a relationship exists between the variables, it is not a linear relationship and so we call it non-linear relationship.

Self-Assessment Exercise 2.

1. Distinguished between linear and non-linear regression.

4.0 Conclusion

You have learned in this unit that the relationship is valid only when it is adequately described by a straight line. You also learned that the equation to this line is called the regression equation and that the equation allows the value of one characteristic to be estimated when the value of other characteristics is known.

5.0 Summary

In this unit the summary of critical concepts learned involve regression analysis which is a statistical technique for predicting the value of one variable, given the value of a second variable.

6.0 Tutor-Marked Assignment

1. Explain the following:

   i. Dependent variable X

   ii. Independent variable

2. Find A if$A = F(\pi r^2)$ and π =22/7, r= 7.
3. Distinguished between linear and non-linear regression.

7.0 Reference/Further Reading

Murray r. Spiegel, Schaum's Outline of Theory and Problems of Statistics (1961), New York, Schaum's Publishing Company.

National Open University of Nigeria, Business Statistics 1 (2006), Sources Plus Ltd, Abuja.

Unit 2

Simple Regression

Content

1.0 Introduction

The general problem of finding equations of approximating curves which fit given set of data is called curve fitting. Individual judgment can often be used to draw an approximate curve to fit a set of data. This is called a freehand method of curve fitting. If the type of equation of this curve is known, it is possible to obtain the constants in the equation by choosing as many points on the curve as there are constants in the equation.

2.0 Learning Outcomes

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

1. Fit a regression line to the given data.

3.0 Learning Content

3.1 The Method of Least Squares

To avoid individual judgment in constructing lines, parabola or other approximating curves to fit set of data, it is necessary to agree on a definition of a 'best fitting line', 'best fitting parabola', etc.

The least square line approximating the set of points (x_1,y_1), (x_2,y_2),…, (x_N,y_N) the equation: $y = a_0 + a_1x$; where the constant $a_0$and a_1 are determined by solving simultaneously the equations.

$$\Sigma y = a_0N + a_1\Sigma x; \quad \Sigma\, xy = a_0\Sigma x + a_1x^2;$$

i.e. $$a_0 = \frac{(\Sigma y)(\Sigma x^2)\text{-}(\ \Sigma x;)\quad(\ \Sigma\, xy)}{(\ \Sigma\, xy)\text{-}\ (\Sigma x^2)}$$

$$a_1 = \frac{\mathrm{N}\Sigma\, xy - (\Sigma x\ )(\Sigma y)}{\mathrm{N}\ (\Sigma x^2)\ \ \text{-}(\Sigma x)^2}$$

The least square parabola approximating the set of the points (x_1,y_1), (x_2,y_2), …, (x_N,y_N) has the equation $y = a_0 + a_1x + a_2x^2$; where constants a_0, a_1 and a_2 are determined by solving simultaneously the equations:

$$\Sigma y = a_0N + a_1\Sigma x + a_2\Sigma x^2$$

$$\Sigma\, xy = a_0\Sigma x + a_1\Sigma x^2 + a_2\Sigma x^3$$

$$\Sigma\, x^2y = a_0\Sigma x^2 + a_1\Sigma x^3 + a_2\Sigma x^4$$

Example

Determine the equation for a straight line which approximates the data:

X	2	3	5	7	9	10
Y	1	3	7	11	15	17

Solution

$$y = a_0 + a_1 x$$

Choose the points (2,1) and (3,3)

For (2,1), $x = 2,\ \ y = 1$

Therefore, $1 = a_0 + 2a_1$ ……………………………………………………………….(1)

For (3,3), $x = 3,\ \ y = 3$

$3 = a_0 + 3a_1$……………………………………………………………….(2)

Solving (1) and (2) simultaneously,

a_0= -3, a_1= 2

The required equation is $y = -3 + 2x$

$$y = 2x - 3$$

It is glaring that the concept behind finding a regression line is based on the assumption that the data points are scattered about a straight line. Howevery in some cases, data points may be scattered about a curve instead of a straight line. In such cases, techniques are available for fitting curves to data points showing a curved pattern. These techniques involve curvilinear regression.

Self-Assessment Exercise

1. A study was conducted so as to predict weights of premature infants based on their ages in weeks. The following observations were collected on 10 infants.

Ages (wks)	6	3	2	2	1	3	4	5	7	4
Weight	6.0	2.5	2.0	2.0	2.0	3.0	4.0	4.5	6.5	3.5

i. Plot the scatter diagram.
ii. Determine the equation of the line relating weight of premature infants to age (in weeks).

4.0 Conclusion

In this unit, a significant and useful measure of association between two characteristics of population was discussed. You have also learned in this unit that the relationship is valid only when it is adequately described by a straight line.

5.0 Summary

In this unit, the summary of critical concepts learned involve regression analysis which is a statistical technique for predicting the value of one variable, given the value of a second variable.

6.0 Tutor-Marked Assignment

1. Explain the following:
   i. Dependent variable
   ii. Independent variable
2. Find A if$A = F(\pi r^2)$ and π =22/7, r= 7.
3. Distinguish between linear and non-linear regression.
4. A study was conducted so as to predict weights of premature infants based on their ages in weeks. The following observations were collected on 10 infants.

Ages (wks)	6	3	2	2	1	3	4	5	7	4
Weight	6.0	2.5	2.0	2.0	2.0	3.0	4.0	4.5	6.5	3.5

i. Plot the scatter diagram.

ii. Determine the equation of the line relating weight of premature infants to age (in weeks).

7.0 References/Further Reading

Murray R. Spiegel, Outline of Theory and Problems of Statistics (1961),Schaum Publishing Company. U.S.A.

National Open University of Nigeria (2004), Introductory to Statistics, Macmillan Publishers Ltd.

Module 7

Correlation Analysis

Unit 1:	Pearson’s Moment of Correlation Coefficient
Unit 2:	Spearman’s Rank Order Correlation Coefficient

Unit 1

Pearson's Moment of Correlation Coefficient

Content

1.0 Introduction

A problem frequently faced in statistics is how to describe a relationship between two or more variables. For instance, is there a relationship between score in a nursing qualifying examination and scores in general examination? You need to determine whether these variables are related.

One of the widely used methods for examining the relationship between two or more variables and for making predictions is correlation analysis.

In this unit we will discuss correlation in term of its meaning, and interpretation.

2.0 Learning Outcomes

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

1. Recognise linear and curvilinear relationship.
2. Display a set of data by means of a scatter diagram.
3. Define correlation.
4. Find Pearson's Moment of Correlation Coefficient (r)

3.0 Learning Content

3.1 Correlation: Meaning and Interpretation

You may have noticed that from the introduction that there are several correlation procedures available. They provide the same type of information on the direction and magnitude of the relationship between the variables.

You should be aware that several correlation procedures are needed because different investigations involved different types of variables with the use of different measuring scales. In addition, you need to be aware that despite the number of different techniques, the same meaning and interpretation are obtained.

3.1.1 Data Arrangement

In a correlation study, data are usually arranged in pairs $(X_{i,}\ Y_i)$. For example, let us see how to represent the following information as data layout for correlation study.

A test considered the measurement of the aptitude of some applicants for admission to nursing school. Estimation of the reliable overtime of the process involved 10 applicants and was given twice within two-week period. The scores for the applicants are now set out in the following layout:

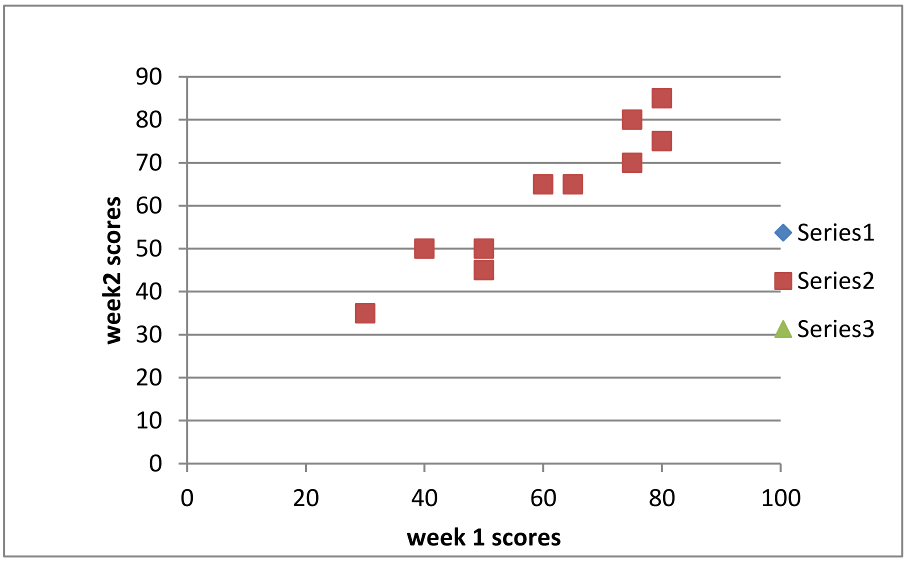
Data layout for correlation study:

	Scores	Scores
Applicant	**Week 1**	**Week 2**
1	80	75
2	40	50
3	75	70
4	50	50
5	30	35
6	50	45
7	75	80
8	60	65
9	65	65
10	80	85

3.1.2. The Scatter Diagram

The initial step in the investigation of a relationship between variables is a graphical display of the data. This display is called scatter diagram and this gives you a visual image of the relationship to be studied.

Do plot each of the n pairs (($X_{i,}$ Y_i) on the graph with the $X'sand\ Y's$ being plotted on the horizontal and vertical axes respectively. The scatter diagram of the data presented in the table above is shown below.



3.3 Pearson's Moment of Correlation Coefficient (r)

Apart from the study of relationships through the scatter diagrams, you need to be aware that a numerical representation of such relationship exists. This is called the correlation coefficient which is the magnitude or strength of the relationship between variables.

The coefficient which measures linear relationship between two continuous variables, is an index number with values ranging from -1 to +1.

You should be aware that a correlation of value 0 is an indicator that there no relationship between the variables, while -1 or +1 is an indicator of either a perfect negative linear relationship or a perfect positive linear relationship. Note that correlation usually lies in the range -1 to +1 with perfect correlation rarely obtained

Product Moment Correlation Coefficient is computed using the following formula:

$$r = \frac{S_{xy}}{\sqrt{\ S_{xx} S_{yy}}}$$

Where $S_{yy} = \Sigma Y^2 - \frac{(\ \Sigma Y)^2}{n}$

$$S_{xx} = \Sigma\ X^2 - \frac{(\Sigma X)^2}{n}$$

$$S_{xy} = \Sigma XY - \frac{(\Sigma X)\ (\Sigma Y)}{n}$$

Example:

A nurse tutor investigated the degree of relationship between student scores on a battery of personality tests ad performance in nursing school and gave a random sample of applicants for nursing school a personality inventory evaluation. The scores on the battery of tests range from 0 to 10. Grade point average in the school was recorded for each student. The data are shown below:

Students	Personality Score	GPA
1	6.0	2.0
2	8.0	3.5
3	7.0	3.0
4	5.0	1.5
5	4.0	1.0
6	3.0	0.5
7	2.0	0.5

ΣXY =75	ΣX^2 = 203 ΣX =35.0	ΣY^2 = 3.0 ΣY =12.0

s TOS_{xx} = $\Sigma X^2 - \frac{(\Sigma X)^2}{n}$

$= 203 - \frac{1225}{7}$

= 203 -175

= 28.0

$S_{yy} = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$

$=30.10 - \frac{144}{7}$

=30.10 – 20.6

= 9.5

$S_{xy} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}$

$= 75 - \frac{42}{7}$

= 75 – 60.7

= 14.3

$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

$= \frac{14.3}{\sqrt{(28.0 \times 9.4)}}$

= 0.88

Self-Assessment Exercise 1

1. Find the coefficient of linear correlation between the variables X and Y presented in the table below:

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

4.0 Conclusion

In this unit, it is noticed that the correlation coefficient is a useful measure of the degree of association between two or more variables. However, this is valid only when a straight line adequately describes this relationship.

5.0 Summary

In this unit, we have seen how to organize bivariate data i.e how to tabulate and diagrammatically represent such data. In particular, we have seen that scatter diagrams are useful to judge what kind of relationship (if any) exists between the two variables. We have also seen that correlation coefficient measures the strength of linear relationship between x and y.

6.0 Tutor Marked Assignment

A statistics tutor investigated the degree of relationship between student scores on a battery of personality tests ad performance in data processing school and gave a random sample of applicants for data processing school a personality inventory evaluation. The scores on the battery of tests range from 0 to 10. Grade point average in the school was recorded for each student. The data are shown below:

Students	**Personality Score**	**GPA**
1	6.0	2.0
2	8.0	3.5
3	7.0	3.0
4	5.0	1.5
5	4.0	1.0
6	3.0	0.5
7	2.0	0.5

a. Co-efficient of the above data.

7.0 Reference/Further Reading

Murray R. Spiegel, Outline of Theory and Problems of Statistics (1961), Schaum Publishing Company. U.S.A.

National Open University of Nigeria (2004), Introductory to Statistics, Macmillan Publishers Ltd.

Unit 2

Spearman’s Rank Order Correlation Coefficient

Content

1.0 Introduction

The spearman's rank order correlation coefficient is a non-parametric or distribution free statistical technique. It is used when the assumptions underlying the classical techniques fail. It is convenient to use this correlation coefficient when data are assigned ranks.

2.0 Learning Outcomes

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

1. Discuss the interpretation of correlation.
2. Discuss the limitations of correlation.
3. Apply the most commonly used correlation procedures.

3.0 Learning Content

Patient	Nurse 1 (X)	Nurse 2 (Y)
1	130	125
2	140	135
3	136	135
4	150	160
5	120	160
6	165	125

3.1 Spearman's Rank Order Correlation Coefficient (r_s)

In a study involving the quality of clinical performance for six students nurses in which evaluation was conducted by two observers, the ranking of student performance was from 1 to 6 by each observer.

Nurse No.	Rank (observer 1)	Rank (Observer 2)
1	2	5
2	3	1
3	4	6
4	5	2
5	1	3
6	6	4

The computation formula for the spearman's rank correlation coefficient is given by

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

With d being the difference between the ranks for each individual and n is the number of pairs of ranks.

$$\Sigma d^2 = (-3)^2 + (2)^2 + (-2)^2 + (3)^2 + (-2)^2 + 2^2$$

$$= 9 + 4 + 4 + 9 + 4 + 4 = 34$$

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6(34)}{6(36 - 1)}$$

$$= 1 - \frac{204}{210}$$

$$= 1 - 0.97$$

$$= 0.03$$

Self-Assessment Exercise 1.

1. The following data represent systolic blood pressure readings (in mmHg) on six patients read by two nurses using the same instruments.

Compute the Spearman’s Rank Order Correlation Coefficient of the data.

3.2 Limitations on the use of Correlation Coefficient

You need to be aware that of the most important rules of interpretation of correlation. These are:

1. Correlation does not mean causation i.e. correlation studies do not prove that one variable causes another. For example, X correlated with Y implies Y correlated with X, so you cannot state that X causes Y, neither can you say that Y causes X.
2. Correlation only applies to the range of values observed for the two variables.
3. Interpretation of correlation is dependent on the particular investigation and judgement of the investigator and the consumer.

Self-Assessment Exercise 2

2. What are the limitations of correlation study?

4.0 Conclusion

In this unit, you have saw that the correlation coefficient is a useful measure of the degree of association between two or more variables but this is valid only when a straight line is adequately describes this relationship.

5.0 Summary

In this unit, the concept of spearman's Rank Order Correlation Coefficient which measures the degree of association between two variables that have been ranked was presented as well as the limitations of correlation studies.

6.0 Tutor-Marked Assignment

1. What are the limitations of correlation study?
2. A statistics tutor investigated the degree of relationship between student scores on a battery of personality tests ad performance in data processing school and gave a random sample of applicants for data processing school a personality inventory evaluation. The scores on the battery of tests range from 0 to 10. Grade point average in the school was recorded for each student. The data are shown below:

Students	Personality Score	GPA
1	6.0	2.0
2	8.0	3.5
3	7.0	3.0
4	5.0	1.5
5	4.0	1.0
6	3.0	0.5
7	2.0	0.5

3. The following data represent systolic blood pressure readings (in mmHg) on six patients read by two nurses using the same instruments.

Patient	Nurse 1 (X)	Nurse 2 (Y)
1	130	125
2	140	135
3	136	135
4	150	160
5	120	160
6	165	125

Compute the Spearman's Rank Order Correlation Coefficient of the data.

7.0 References/Further Reading

Murry R. Spiegel, Outline of Theory and Problems of Statistics (1961),Schaum Publishing Company. U.S.A.

National Open University of Nigeria (2004), Introductory to Statistics, Macmillan Publishers Ltd.

Module 8

Index Numbers

Unit 1

Index Numbers

Content

1.0 Introduction

An index number is a statistical measure designed to show case changes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc. A collection of index numbers for different years, locations, etc., is sometimes called an index series.

2.0 Learning Outcomes

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

1. Define Index Numbers.
2. Mention different index numbers we have.
3. Obtain index number of a given set of data.

3.0 Learning Content

3.1 Index Numbers

By using index numbers, we can, for example, compare food or other living costs in a city during one year with those of previous year, or we can compare steel production during a given year in one part of a country with that in another part. Although mainly used in business and economics, index numbers can be applied in many other fields.

Many governmental and private agencies are engaged in computation of index numbers or indexes, as they are often called, for the purposes of forecasting business and economic conditions, providing general information, etc. Thus, we have wage indexes, production indexes, unemployment indexes, and many others. Perhaps the most well-known is the cost of living index or consumer price index.

Self-Assessment Exercise(s) 1

1. What do you understand by Index Numbers
2. What are the applications of Index Numbers

3.2 Price Relatives

One of the simplest examples of index number is a price relative, which is the ratio of the price of a single commodity in a given period to its price in another period called the base period or reference period. For simplicity we assume price to be constant for one period. If they are not, an approximate average for the period can be taken to make this assumption valid.

If P_o and P_n denote the commodity price during the base period and given period respectively, then by definition

$$\text{Price relative} = \frac{P_n}{P_o} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(1)}$$

And is generally, expressed as a percentage by multiplying by 100

More generally if $P_0$and P_n are prices of a commodity during periods a and b respectively, the price relative in period b with respect to period a is defined as $P_b\,/P_a$ and is denoted by $P_{a/\mathrm{b}}$, a notation which will be found useful. With this notation the price relative in equation (1) can be denoted by $P_{o/\mathrm{n}}$.

Example

Suppose the prices of a quart of milk in the years 1995 and 2005 were ₦25.00 and ₦30.00 respectively. Calculate the price relative taken 1995 as the base year and interpret your result.

Solution:

Price relative = $P_{1995/2005}$

$$= \frac{\text{price in 2005}}{\text{Price in 1995}}$$

$$= \frac{₦30.00}{₦25.00}$$

= 1.2

=120%

This result simply means that in 2005 the price of milk was 120% of that in 1995, i.e it increased by 20%.

Self-Assessment Exercise(s) 2

1. Explain briefly the term price relative.
2. Suppose the prices of a quart of milk in the years 1995 and 2005 were ₦25.00 and ₦40.00, respectively. Calculate the price relative taken 2005 as the base year and interpret your result.

3.3 Quantity or Volume Relatives

Instead of comparing prices of a commodity, we may be interested in comparing quantities or volumes of the commodity, such as quantity or volume of production, consumption, exports, etc. in such cases we speak of quantity relatives or volume relatives. For simplicity, as in the case of prices, we assume that quantities are constant for any period. If they are not, an appropriate average for the period can be taken to make this assumption valid.

If q_0 denotes the quantity or volume of a commodity produced, consumed, exported, etc., during the base period, while q_n denotes the corresponding quantity produced, consumed, etc., during a given period, we defined

Quantity or volume relative $= \dfrac{q_n}{q_0}$ ……………………….(2)

which is generally expressed as a percentage.

As in the case of price relatives, we use the notation $q_{a/\mathrm{b}} = q_b / q_a$ to denote the quantity in period b with respect to period a. The same remarks and properties pertaining relatives are applicable to quantity relatives.

Self-Assessment Exercise(s) 3

1. Explain the term Quantity or Volume Relatives.
2. Differentiate between price relative and quantity or volume relatives

3.4 Simple Aggregate Method of Computing Price Index

In this method of computing a price index, we express the total of commodity prices in the given year as a percentage of total commodity prices in the base year. In symbols, we have

Simple aggre**gate price** index $= \dfrac{\Sigma\, P_n}{\Sigma\, P_o}$ ……………………………………….(3)

Where $\Sigma\, P_o$ = sum of all commodity prices in the base year.

$\Sigma\, P_n$ = sum of corresponding- commodity prices in the given year

And where the result is expressed as a percentage, as are index numbers in general.

Example:

The table below shows Nigeria wholesale prices and production of fluid milk, butter and cheese for the years 2000, 2001 and 2002. Compute a simple aggregate wholesale price index of these dairy products for the year 2002 using i. 2000 and ii.2000-2001 as a base.

Solution:

i. Simple aggregate price index = $\frac{\Sigma P_n}{\Sigma P_o}$

$= \frac{\text{sum of prices in the given year 2002}}{\text{Sum of prices in the base year 2000}}$

$= \frac{4.13 + 59.7 + 38.9}{39.5 + 61.5 + 34.8}$

$= 101.8\ (\%)$

i.e average whole prices in 2002 are 102.5% of those in 2000 (or 2.5% greater).

ii. Average (mean) price of milk in base period 2000-2002 = $\frac{1}{2}$(3.95 + 3.89)
=₦3.92k/1 b

Average (mean) price of butter in base period 2000-2002 = $\frac{1}{2}$(61.5 + 62.2)
=₦61.85k/1 b

Average (mean) price of cheese in base period 2000-2002 = $\frac{1}{2}$(34.8 + 35.4)
=₦35.1k/1 b

Simple aggregate price index = $\frac{\Sigma P_n}{\Sigma P_o}$

$= \frac{\text{Sum of prices in the given year 2002}}{\text{Sum of prices in the base period 2000-2002}}$

$= \frac{4.13 + 59.7 + 38.9}{3.92 + 61.85 + 35.1}$

$= 101.8(\%)$

Self-Assessment Exercise 4

1. The table below shows Nigeria retail prices and production of coal and gasoline for the years 1960 and 1965. Compute a simple aggregate wholesale price index of these products for the year 1965 using 1960 as a base.

	Price p in (₦) & Quantity produced q	Price p in (₦) & Quantity produced q
	1960	1965
coal **p**	20.13 per short ton 3.559 million short per tons	28.20 per short ton 1.821 million short per tons
gasoline **p** **q**	20.3 per short ton 80.2 million barrels •	21.4 per short ton 118.6 million barrels •

• each barrel contains 42 gallons.

4.0 Conclusion

In this unit, we have been able to show that an index number is a statistical measure designed to show case changes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc. A collection of index numbers for different years, locations, etc., is sometimes called an index series.

5.0 Summary

In summary, by using index numbers we can, for example, compare food or other living costs in a city during one year with those of previous year, or we can compare steel production during a given year in one part of a country with that in another part. Although mainly used in business and economics, index numbers can be applied in many other fields.

Many governmental and private agencies are engage in computation of index numbers or indexes, as they are often called, for the purposes of forecasting business and economic conditions, providing general information, etc. Thus, we have wage indexes, production indexes, unemployment indexes, and many others. Perhaps the most well-known is the cost of living index or consumer price index.

6.0 Tutor-Marked Assignment

1. What do you understand by Index Numbers
2. What are the applications of Index Numbers

3. The table below shows Nigeria retail prices and production of coal and gasoline for the years 1960 and 1965. Compute a simple aggregate wholesale price index of these products for the year 1965 using 1960 as a base.

	Price p in (₦) & Quantity produced q	Price p in (₦) & Quantity produced q
	I960	1965
coal p q	20.13 per short ton 3.559 million short per tons	28.20 per short ton 1.821 million short per tons
gasoline p q	20.3 per short ton 80.2 million barrels •	21.4 per short ton million barrel •

• each barrel contains 42 gallons.

7.0 References/Further Reading

Murry R. Spiegel, Outline of Theory and Problems of Statistics (1961),Schaum Publishing Company. U.S.A.

National Open University of Nigeria (2004), Introductory to Statistics, Macmillan Publishers Ltd.

Answers to Self-Assessment Exercises

Module One: Nature and Sources of Statistics Data

Unit 1: Nature of Statistical Data

SAEs 1

1. This refers to numerical description of quantitative aspect of things and this description may take the form of courts or measurement.
2. Statistical method is a technique used to obtain analyses or present numerical data. Examples of statistical techniques include:
    i. Collection and assembling of data.
    ii. The classification and condensation of data.
    iii. Presentation of data in either tabular form or graphical form.
    iv. Analysis of data.

SAEs 2

Statistics can be defined as a branch of science that deals with collection, organization, classification, analysis, interpretation, presentation of data expressed in numerical form for the purpose of drawing valid conclusions.

i. The descriptive statistics is concerned with processing, summarizing conclusions and presentation of data. This could be in the form of tables or graphs to reveal some inherent information contained in a data set; and to present the information in convenient form.
ii. The inferential statistics utilizes sample data to make estimates, decisions, prediction or generalization about a population.
iii. The inductive or analytical statistics is concerned with the method of analysis. Statistics enables us to be able to evaluate data sets intelligently.

Unit 2: Importance of Statistics

SAE 1

Importance of Statistics are to:

i. Present a large amount of quantitative information
ii. In an organised way, go beyond a meaningless collection of data to a meaningful interpretation
iii. Predict how likely an event will occur
iv. Make inferences from observations
v. Save us time and energy by condensing large amount of information concisely and conveniently in a table.

Unit 3: Sources of Statistical Data

SAEs 1

1. A population is the set of all units or objects in a defined area of interest e.g. HIV patients in Niger State.
2. A sample is a subset or fraction of the units of the population of interest.
3. i. Analysis based on sampling is as precise as that based on the entire population.
   ii. Use of sampling is time saving.
   iii. Sampling is cost minimizing-both human and material costs.
   iv. Analysis based on sample is of greater accuracy than that based on the entire population.
   v. The use of population to obtain some of its parameters may not be feasible (practicable)

SAE 2

i. Published Sources
ii. Sample survey
iii. Designed Experiments
iv. Direct Observation

Unit 4: Methods of Data Collection

SAEs 1

1. i. They must in aggregate.

ii. They must be enumerated to a marked extent by a multiplicity of causes.

iii. The y must be enumerated or estimated according to reasonable standard accuracy.

iv. They have been collected in a systematic manner for a predetermined purpose

v. They must be comparable

2. i. Primary data

ii. Secondary data.

3. i. Qualitative data

ii. Quantitative data

SAEs 2

1. i. Interview
ii. Questionnaire
iii. Transcription from the records.
iv. Direct observations.

2. This is a sampling procedure which consists of stratifying (dividing the population into a number of non —overlapping sub population called strata), then testing a

sampling from every stratum. The items or sample from each strum can then be selected by any suitable random method. Stratified sampling procedure is very good and appropriate when our population is large and heterogeneous.

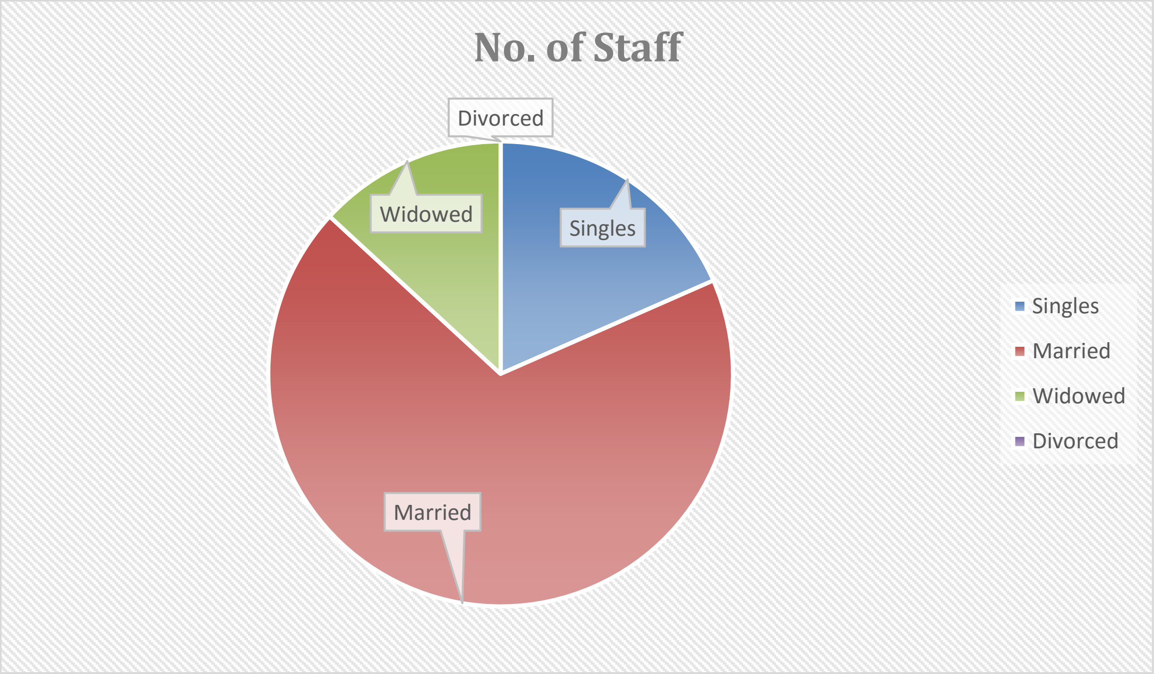
Module Two: Data Presentation

Unit 1: Data Presentation

SAE 1

1. Yes

SAE 2



Total no. of staff in the institution is
$3 + 130 + 25 + 10 = 200$
Angles corresponding to each status are found thus:

Single: $= \frac{35}{200} \times 360° = 63°$

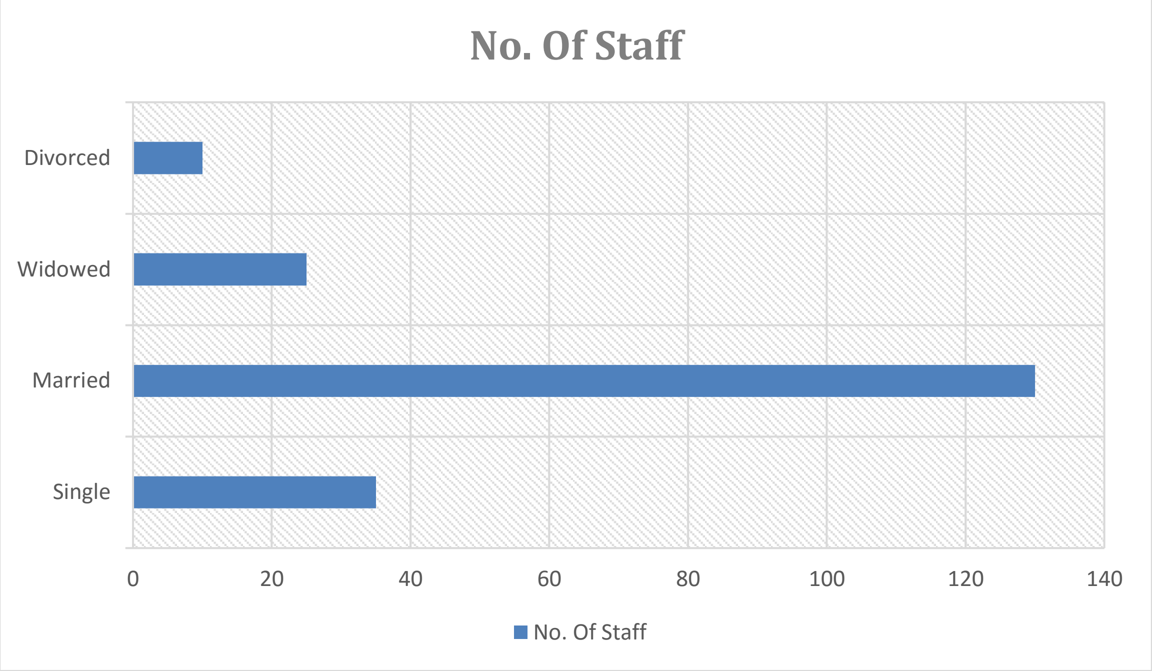
Married: $= \frac{130}{200} \times 360° = 234°$

Widowed:$= \frac{25}{200} \times 360° = 45°$

Divorced: $= \frac{10}{200} \times 360° = 200°$

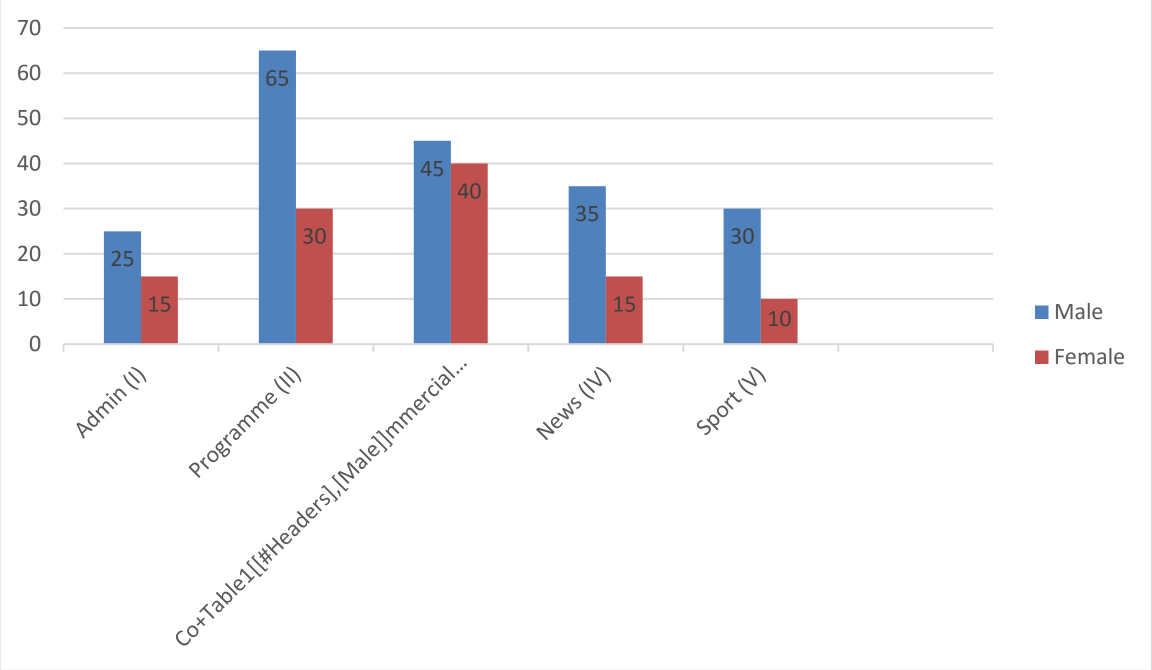
Thus, the pie chart is:

SAE 3

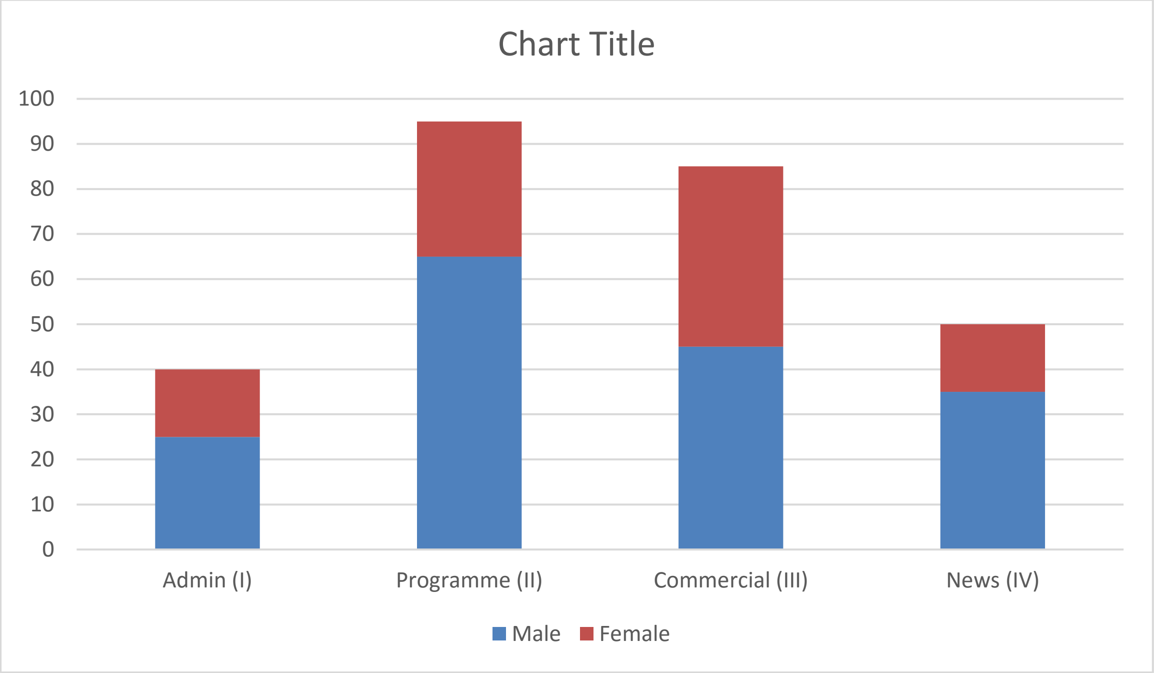


SAE 4

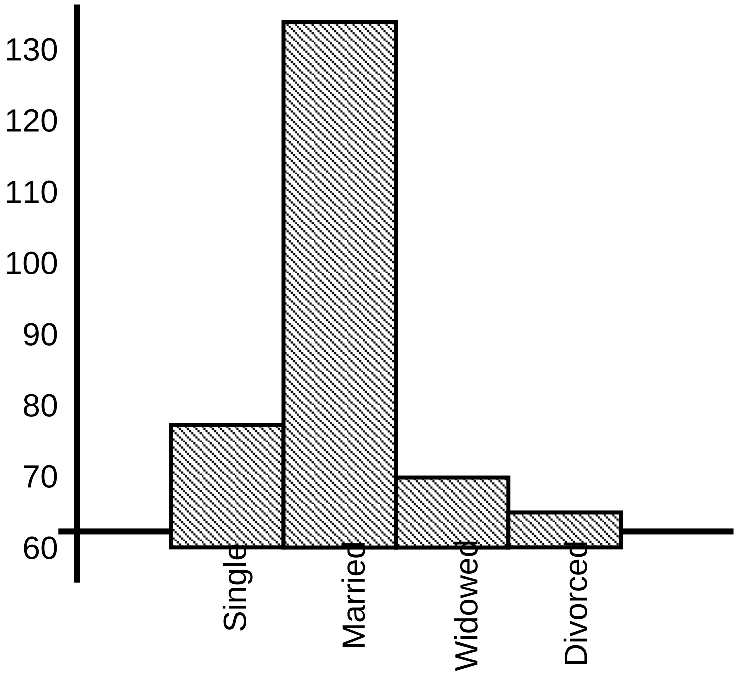
i. Multiple Bar Chart.



ii. Compound Bar Chart.

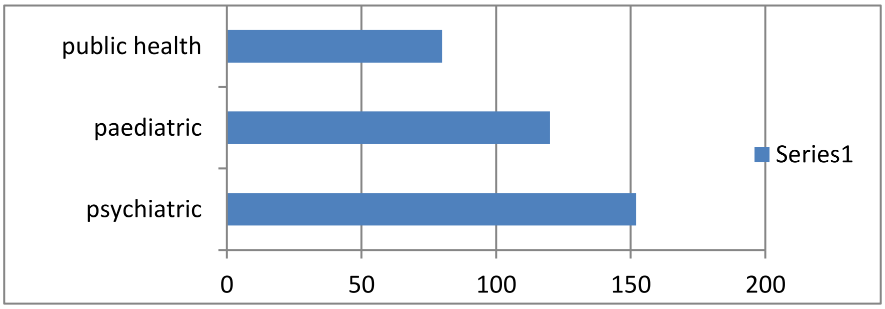


SAE 5

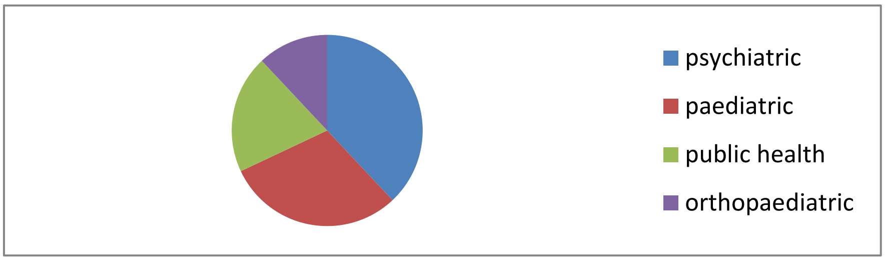


SAE 6

1.



2. .



Module Three: Measures of Central Tendency

Unit 1: Measure of Location

SAE 1

Weights (kg)	Tally marks	Frequency
48 – 52	IIII III	8
53 – 57	IIII IIII II	12
58 – 62	IIII IIII	10
63 – 67	IIII I	6
68 – 72	IIII	4
Total		40

SAE 2

SAE 3

Weights (kg)	f	X	fx	D = x – A = x – 61	Fd	u	fu
48 – 52	8	50	400	-11	-88	-2	**8**
53 – 57	12	55	660	-6	-72	-1	**6**
58 – 62	10	60	600	-1	-10	0	**0**
63 – 67	6	65	390	4	24	1	**6**
68 – 72	4	70	280	9	36	2	**8**
Total	**40**		**2330**		**-110**		**-14**

a. **Long method**

$$x = \frac{\Sigma f x}{\Sigma f}$$

$$=\frac{2330}{40}$$

$$= 58.25$$

b. Assumed Mean of 61:

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f}$$

$$\bar{x} = 61 + \frac{-110}{40}$$

$$= 61 - 2.75 = 53.25$$

c. Coding method:

$$\bar{x} = A + \left(\frac{\Sigma fu}{\Sigma f}\right) + C$$

Here C = 5. A is the value of x corresponding to u = 0

For odd number of classes we choose u = 0 at the centre

Thus, $\bar{x} = 60 + \left(\frac{-14}{40}\right) + 5$

= 60 + 1.75

= 58.25

SAEs 4

Merits of Mean

It takes account of all the values of a distribution. It is therefore, more representative than the other two and for this reason alone, it is used more than the other two averages.

Demerits of Mean

1. It is often the most difficult to calculate.
2. It is not easily understood by non-statistician.
3. While the mice and median often represent actual scores belonging to some members of the population, the. Mean often does not.
4. When the mean is used with discrete variables (e.g. number of children), it often yields unrealistic values such as 2.5 children.
5. While the median and mode can be obtained graphically the mean cannot.

SAEs 5

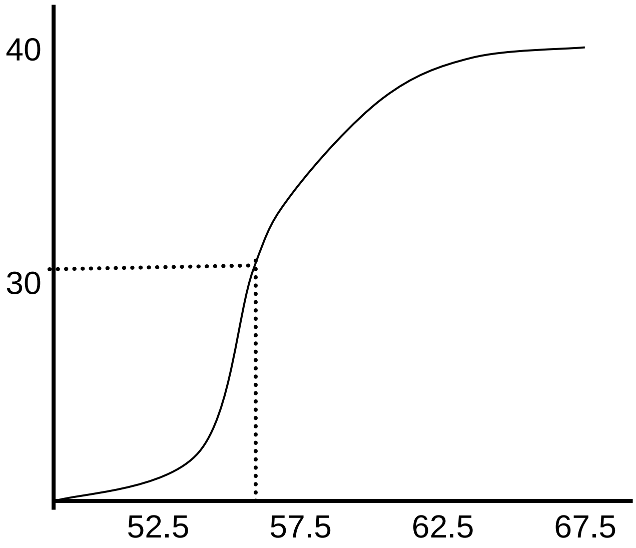
1. Merits of Median

1. It is easily understood.
2. It is relatively easy to calculate.

Demerits of Median

1. It takes no account of extreme values in the distribution. For instance, the median of 2, 40, 43, 45, and 96 is even though there are two extreme values 2 and 96.
2. It does not use all the data available.

2.



Estimated median = 57.5

b. $$\text{Median} = L + \left(\frac{N/2 - F}{f}\right) C$$

Now, N/2 = 40/2 = 20

I.e. the median is the 20th value. From the cumulative frequency distribution table, 20th item falls within the class 53 – 57. Thus, the median class is 53 – 57. Hence:

$$\text{Median} = 52.5 + \left(\frac{20 - 8}{12}\right) \times 5$$

$$= 52.5 + 5 \quad = 57.5$$

Comparison: The graphical and the calculated values agree appreciably, depending on the accuracy of your graph.

SAE 6

Class interval	**F**	X	Fx	D = x-40.5	Fd	U	Fu
60-**69**	5	64.5	322.5	24	120	3	15
50-59	11	54.5	599.5	14	154	2	22

40-49	17	45.5	773.5	5	85	1	17
30-39	14	35.5	497	-5	-70	0	0
20-29	9	25.5	229.5	-15	-135	-1	-9
10-19	4	15.5	62	-25	-100	-2	-8

i. $\bar{x} = \frac{\Sigma \mathrm{fx}}{\Sigma f}$

$= \frac{2484}{60}$

$= 41.4$

ii. $\bar{x} = \mathrm{A}+ \frac{\Sigma \mathrm{fd}}{\Sigma f}$

$= 40.5 + \frac{54}{60}$

$= 41.4$

$\bar{x} = \mathrm{A}+ \frac{\Sigma \mathrm{fu}}{\Sigma f}$

$= 35.5+ \frac{37}{60}$

$= 41.4$

Unit 2: Weighted Mean

SAEs 1

1. Yes
2. $\bar{X} = \frac{1(70) + 1(90) + 3(85)}{1 + 1 + 3}$

$= \frac{415}{5}$

= 83

SAEs 2

1. Yes

2. GM = $\sqrt[4]{}(3x5x6x7)$

a. $= \sqrt[4]{630}$

b. =5.01

3. HM = $\underline{\ 3}$

c. $\frac{1}{5}+\frac{1}{6}+\frac{1}{7}$

d. = $\underline{3}$

e. 0.2+0.0167+0.143

f. = $\underline{3}$

g. 0.5099

= 5.88

SAE 3

1. The geometric mean of a set of positive numbers $X_1, X_2, ..., X_r$ is less than or equal to their arithmetic mean but is greater than or equal to their harmonic mean. In symbols,

$$H \leq G \leq X$$

The equality signs hold only if all the numbers $X_1, X_2, . . . , X$ are identical.

Module Four: Measures of Dispersion

Unit 1: Measure of Dispersion 1

SAE 1

1. R = 93- 34

= 59

SAE 2

Q_1 = ₦59.995 + $\frac{8.25}{10}$(₦10)

= ₦68.25

Q_3 = ₦89.995 + $\frac{0.75}{10}$(₦10)

= ₦90.75

SAEs 3

1. $D_2 = ₦59.995 + \frac{5}{10}(₦10)$

$= ₦65$

$D_4 = ₦69.995 + \frac{8}{16}(₦10)$

$= ₦75$

2. Percentiles are ordinal measures; they are scores points which divide the distribution into 100 equal parts called percentages

Unit 2: Measure of Dispersion 2

SAE 1

1. 4.25

SAE 2

$S=\sqrt{\frac{873}{100}} - (\frac{45}{100})^2$

$= \sqrt{8.5275}$

$=2.92$

$S^{2 =} 8.53$

Module Five: Skewness and Kurtosis

Unit 1: Skewness and Kurtosis

SAE 1

1. Mean = ₦79.76
   Median = ₦79.06
   Mode = ₦77.50
   Standard deviation = S
   =₦15.60

   Coefficient of skewness $= \frac{3(\text{mean - median})}{S}$

   $= \frac{3(₦79.76 - ₦79.06)}{₦15.60}$

   $= 0.1346$ or 0.13

   Since the coefficient is positive, the distribution is skewed positively, i.e to the right.

SAE 2

1. a. The degree of skewness is the extent to which the given distribution departs from symmetry.
   b. It is the degree of steepness or pointedness of distribution.

Module Six: Simple Regression

Unit 1: Dependent and Independent Variables

SAE 1

1. If to each value which a variable X can assume, there are corresponds one or more values of variable Y, we say that Y is a function of X and write Y=F(x) (read ‘Y equals F of X’). To indicate this functional dependence, the variable X is called the independent variable and Y is called the dependent variable
2. $A = F(\pi r^2)$

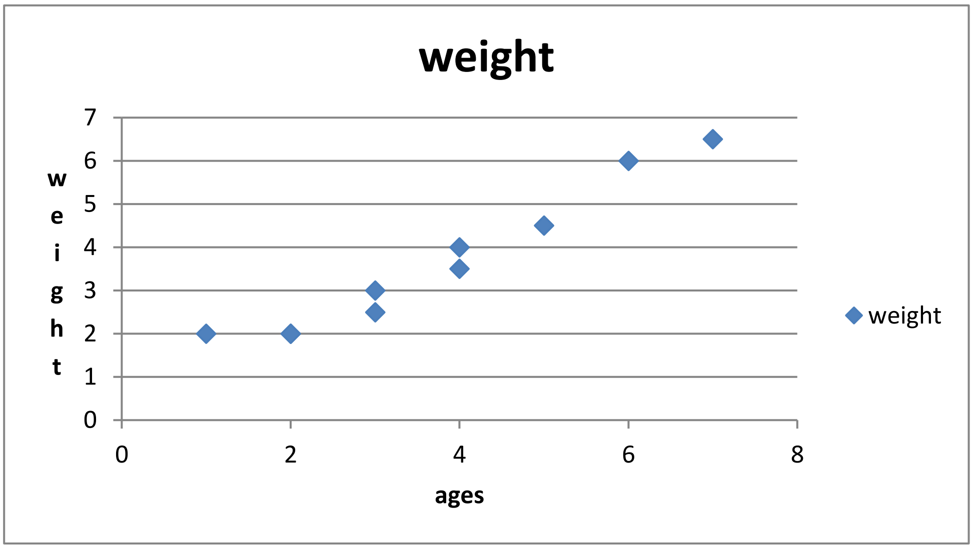
$$= \frac{Fx22}{7}\ x7x7$$

A= 154F

Unit 2: Simple Regression

SAE 1

1. The scatter diagram



$$\bar{y} = 3.6$$

2. $\bar{x} = 3.7$

$s_{yy} = 14.4$

$s_{xx} = 32.1$

$s_{xy} = 27.3$

β =$\frac{27.3}{32.1} = 0.85$

α =3.6 – $\frac{27.3}{32.1}$ (3.7)

= 3.6 – 0.85(3.7)

= 3.6 – 3.15

= 0.450

$y = 0.45 + 0.85x$

Module Seven: Correlation Analysis

Unit 1: Pearson's Moment of Correlation Coefficient

SAE 1

1. $S_x = 4.06$
   $S_y = 2.65$
   $S_{xy} = 10.50$
   $r = \frac{S_{xy}}{S_x S_y}$

   =$\frac{10.50}{(4.06)(2.65)} = 0.976$

Unit 2: Spearman's Rank Order Correlation Coefficient

SAE 1

1. r_s = 1 – $\frac{6\Sigma d^2}{n(n^2 \; - 1)}$

$\Sigma d^2 = 41.50$

$n = 6$

$r_s = 1 - 6(41.50)/6(36 - 1)$

= 1−1.1857

=−0.19

SAE 2

1. Correlation does not mean causation i.e. correlation studies do not prove that one variable causes another.
2. Correlation only applies to the range of values observed for the two variables.
3. Interpretation of correlation is dependent on the particular investigation and judgement of the investigator and the consumer

Module Eight: Index Numbers

Unit 1: Index Numbers

SAEs 1

1. An index number is a statistical measure designed to show case changes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc
2. i. For purpose of forecasting business and economic conditions.

   ii. For the purpose of comparison

SAEs 2

1. Price relative is the ratio of the price of a single commodity in a given period to its price in another period called the base period or reference period.
2. Price relative = $P_{1995/2005}$

$$= \frac{\text{price in 1995}}{\text{Price in 2005}}$$

$$= \frac{₦25}{₦40}$$

= 0.65

=65%

SAEs 3

1. Instead of comparing prices of a commodity, quantity or volume relatives compares quantities or volumes of the commodity, such as quantity or volume of production, consumption, exports, etc. in such cases we speak of quantity relatives or volume relatives.
2. Price relatives deal with ratio of price of a single commodity in a given year to its price in another period called base period or reference period while Quantity or Volume relatives compares quantity or volume of the commodity instead of prices.

SAE 4

$$\frac{\Sigma P_n}{\Sigma P_O} = sum\ of\ prices\ in\ a\ given\ year\ (1965)$$

$$/sum\ of\ prices\ in\ a\ base\ year(\ 1960)$$

$$\frac{= ₦28.20 + ₦0.214}{₦20.13 + ₦0.203} = 139.\%$$

Indicating that the average retail prices of those commodities in 1965 were 37.9% greater than in 1960.

DAYSPRINGERS PRODUCTION
+2347086257747