# COURSE GUIDE

**FSS 211**
**SOCIAL SCIENCE STATISTICS**

**Course Team** Dr. Emmanuel Ifeanyi AJUDUA & Vivian Anietem ODISHIKA-(Course Developers/Writers)-NOUN
Professor Anthony AKAMOBI (Course Editor) - Chukwuemeka Odumegwu Ojukwu University Anambra State


**NATIONAL OPEN UNIVERSITY OF NIGERIA**

National Open University of Nigeria
Headquarters
University Village
Plot 91, Cadastral Zone
Nnamdi Azikiwe Expressway
Jabi, Abuja

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

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



Printed 2020

ISBN: 978-978-970-152-0

**CONTENT** **PAGE**

## INTRODUCTION

Welcome to FSS211 SOCIAL SCIENCE STATISTICS.
FSS211: Social Science Statistics is a two-credit and one-semester undergraduate course for social science students. The course is made up of ten units spread across fifteen lecture-weeks. This course guide gives you an insight to Social Science Statistics in a broader way and how to make use and apply statistics as a social scientist. It tells you about the course materials and how you can work your way through these materials. It suggests some general guidelines for the amount of time required of you on each unit in order to achieve the course aims and objectives successfully. Answers to your Tutor-Marked Assignments (TMAs) are therein already.

This course is basically on Social Science Statistics. As a social scientist, it is expected that you should be able to apply statistical tools and techniques to social science problems. The topics covered include Meaning, Types, Concepts and Importance of Statistics, Meaning, Types and Classification of Statistical Symbols, Variables and their Measurements, Measures of Central Tendency and Dispersion, Data Collection and Presentation, Probability, Sampling, Hypothesis and Significance Testing, Correlation and Regression analysis.

## COURSE AIMS

The aim is to help you have the basic knowledge of statistics as it relates to research in social sciences. However, the following broad aims will also be achieved:

- Help students understand the importance of statistics and statistical symbols
- Acquaint students with the concept of variables in statistics and its role in statistical analysis
- To familiarise students with the knowledge of hypotheses testing
- Help students understand some basic statistical applications with regards probability and measures of central tendencies and dispersions.
- Stimulate student’s knowledge on sampling
- To ensure students understand different types of data and how to collect them
- To make the students to understand and apply statistical calculation such as t test, f test and chi-square analysis.
- To expose the students to analysis of simple linear regression analysis

- To ensure that the students know how to apply simple linear regression to economics situations.
- To make the students to be to interpret simple linear regression analysis result based on findings.

## COURSE OBJECTIVES

Generally, the objective of FSS 211 is centred on equipping social science students with necessary statistical knowledge. This will be of great use to you as a social scientist. Each unit in the course material has its own objective(s) which has been clearly stated at the beginning of each unit. It is advisable that you read them before working through the units. References may be made to them in the course of studying the units. On the successful completion of the course, you should be able to:

a. Discuss the definitions, meaning, types and importance of statistics.
b. Elucidate the several approaches to data collection, their merits and limitations.
c. Discuss variables, types, measurement and their relationship.
d. Understand the classification of data.
e. Present data using tables, graphs and charts.
f. Analyze measures of central tendencies and dispersion and their usefulness in statistics.
g. Explain the probability theory.
h. Understand correlation and its application.
i. Understand the nature, meaning and importance of regression analysis and its application.

## WORKING THROUGH THE COURSE

To successfully complete this course, you are required to read the study units, referenced books and other materials on the course.
Each unit contains self-assessment exercises called Student Assessment Exercises (SAE). At some points in the course, you will be required to submit assignments for assessment purposes. At the end of the course there is a final examination. This course should take about 15weeks to complete and some components of the course are outlined under the course material subsection.

## COURSE MATERIAL

The major component of the course, what you have to do and how you should allocate your time to each unit in order to complete the course successfully on time are listed as follows:

1. Course guide
2. Study unit
3. Textbook
4. Assignment file
5. Presentation schedule

## STUDY UNIT

There are ten (10) units in this course which should be studied carefully and diligently.

**Module 1 Introduction, Concepts and Methods of Statistics**

Unit 1 Meaning, Types, Concepts and Importance of Statistics
Unit 2 Meaning Types and Classification of Statistical Symbols
Unit 3 Variables and their Measurement

**Module 2 Descriptive Statistics and Probability**

Unit 1 Measure of Central Tendency and Dispersion
Unit 2 Data Collection and Presentation
Unit 3 Probability

**Module 3 Inferential Statistics**

Unit 1 Sampling
Unit 2 Hypothesis and Significance Testing
Unit 3 Correlation
Unit 4 Regression Analysis

Each study unit will take at least two hours, and it include the introduction, objective, main content, self-assessment exercise, conclusion, summary and reference. Other areas border on the Tutor-Marked Assessment (TMA) questions. Some of the self-assessment exercise will necessitate discussion, brainstorming and argument with some of your colleagues. You are advised to do so in order to understand and get acquainted with statistics and its application.

There are also textbooks under the references and other (on-line and off-line) resources for further reading. They are meant to give you

additional information if only you can lay your hands on any of them. You are required to study the materials; practice the self-assessment exercise and tutor-marked assignment (TMA) questions for greater and in-depth understanding of the course. By doing so, the stated learning objectives of the course would have been achieved.

## TEXTBOOKS AND REFERENCES

Adedayo, A.O. (2006). *Understanding Statistics.* Lagos: JAS Publishers.

Anderson, D. R., Sweeney, D. J. & Williams T. A. (2008). *Statistics* for Business and Economics. (10th ed.). USA: Thomson Corporation.

Bobko, P. (2001). *Correlation and Regression: Applications for Industrial Organisational Psychology and Management* (2nd ed.). Thousand Oaks, CA: Sage Publications

Carlson, R. (2006): *A Concrete Introduction to Real Analysis*. CRC Press.

Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and Nonparametric Measures.* Thousand Oaks, CA: Sage Publications.

Cramer D. & Howitt D. (2004). *The SAGE Dictionary of Statistics a Practical Resource for Students in the Social Sciences.* London*:* SAGE Publications.

Damodar, N. G., Dawn, C. P., & Sangetha, G. (2012). *Basic Econometrics.* New Delhi: Tata McGraw Hill Education Private Ltd.

Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms*, OUP. ISBN 0-19-920613-9

Dominick, S. & Derrick, R. (2011). *Statistics and Econometrics (Schaum Outlines)*. NewYork: McGraw-Hill Company,

Esan F.O. and Okafor, R.O. (2010): *Basis Statistical Methods (revised edition)* Lago: Toniichristo Concept

Everitt, B. S. & Skrondal A. (2010). *The Cambridge Dictionary of Statistics.*

Everitt, B. S. (2002). *The Cambridge Dictionary of Statistics.* (2nd ed.). Cambridge UP. ISBN 0-521-81099-X.

Grewal, P. S. (1990). *Methods of Statistical Analysis.* (2nd ed.). New Delhi: Sterling Publishers pvt. Ltd.,

Grinstead& Snell's (2006). *Introduction to Probability*. The CHANCE Project1, Version dated 4 July 2006.

Gupta, C. B. (1983). *An Introduction to Statistical Methods.* New Delhi: Vikas Publishing House PVT Ltd.

Kozak, A., Kozak, R., Watts, S., & Staudhammer, C. (2008). *Introductory Probability and Statistics.* Vancouver: Cabi International.

Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimmer, H., (2014). *Introduction to Statistics*. Rice University, Houston, TX. Accessed: http://onlinestatbook.com/Online_Statistics_Education.pdf.

Levine, D. M. & Stephan, D. F. (2005). *Even You Can Learn Statistics. A Guide for Everyone Who Has Ever Been Afraid of Statistics*. Pearson Education, Inc. Publishing as Pearson Prentice Hall Upper Saddle River, NJ. USA Accessed: file:///C:/Users/1/Downloads/[David_M._Levine,_David_F._Stephan]_Even_you_can_l(z-lib.org)%20(1).pdf

National Bureau of Statistics (2018). *National Bureau of Statistics 2017 Demographic Statistics:* Bulletin. https://nigerianstat.gov.ng/download/775.

Obikeze, D. S (1986). *Introductory Statistics for the Social Sciences.* Enugu: Fourth Dimension Publishing Co. Ltd.

Okoro, E. (2002). *Quantitative Techniques in Urban Analysis.* Ibadan: Kraft Books Ltd.

*Webster's Revised Unabridged Dictionary. G & C Merriam, 1913*,Publishing Co. Ltd, Probability.

Shafer, D. S. & Zhang, Z. (2012). *Introductory Statistics*. Flat World Knowledge. Washington, DC, USA.

Tokunaga H. T. (2016). *Fundamental Statistics for the Social and Behavioral Sciences*. California: SAGE Publications.

Watkins, J. C. (2016).*An Introduction to the Science of Statistics*: From Theory to Implementation (Preliminary Edition). Accessed: https://www.math.arizona.edu/~jwatkins/statbook.pdf.

## ASSIGNMENT FILE

Assignment files and marking scheme will be made available to you. This file presents you with details of the work you must submit to your tutor for marking. The marks you obtain from these assignments shall form part of your final mark for this course. Additional information on assignments will be found in the assignment file and later in this Course Guide in the section on assessment.

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

Assignment 1 - All TMAs' question in Units 1 – 3 (Module 1)
Assignment 2 - All TMAs' question in Units 4 – 6 (Module 2)
Assignment 3 - All TMAs' question in Units 7 – 10 (Module 3)

## ASSESSMENT

There are two types of the assessment for the course. First is the tutor-marked assignment; second, is the examination. In attempting the assignments, you are expected to apply information, knowledge and techniques gathered during the course. The assignments must be submitted to your tutor for formal Assessment in accordance with the deadlines stated in the Presentation Schedule and the Assignments File. The work you submit to your tutor for assessment will count for 30% of your total course mark. At the end of the course, you will need to sit for a final examination of two hour's duration. This examination will account for 70% of your total course mark.

## TUTOR-MARKED ASSIGNMENTS (TMAS)

There are three tutor-marked assignments in this course. You will submit all the assignments. You are encouraged to work on all the questions thoroughly. The TMAs constitute 30% of the total score. Assignment questions for the units in this course are contained in the Assignment File. You will be able to complete your assignments from the information and materials contained in your set books, reading and study units. However, it is desirable that you demonstrate that you have read and researched more widely than the required minimum. You should use other references to have a broad viewpoint of the subject and also to give you a deeper understanding of the subject.

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

## FINAL EXAMINATION AND GRADING

The final examination will be of two hours' duration and have a value of 70% of the total course grade. The examination will consist of questions which reflect the types of self-assessment practice exercises and tutor-marked problems you have previously encountered. All areas of the course will be assessed.

Revise the entire course material using the time between finishing the last unit in the module and that of sitting for the final examination. You might find it useful to review your self-assessment exercises, tutor-marked assignments and comments on them before the examination. The final examination covers information from all parts of the course.

## COURSE MARKING SCHEME

The Table presented below indicates the total marks (100%) allocation.

| **Assignment** | **Marks** |
|---|---|
| Assignments (Three assignments submitted and marked) | 30% |
| Final Examination | 70% |
| **Total** | **100%** |

## COURSE OVERVIEW

The Table presented below indicates the units, number of weeks and assignments to be taken by you to successfully complete the course, Social Science Statistics (FSS211).

| **Units** | **Title of Work** | **Week's Activities** | **Assessment (end of unit)** |
|---|---|---|---|
| | Course Guide | | |
| **MODULE 1: INTRODUCTION, CONCEPTS AND METHODS OF STATISTICS** | | | |
| 1 | Meaning, Types, Concepts and Importance of Statistics | Week 1 | Assignment 1 |
| 2 | Meaning Types and | Week 2 | Assignment |

| | | | |
|---|---|---|---|
| | Classification of Statistical Symbols | | 1 |
| 3 | Variables and their Measurement | Week 3 | Assignment 1 |
| **MODULE 2: DESCRIPTIVE STATISTICS AND PROBABILITY** | | | |
| 1 | Measure of Central Tendency and Dispersion | Week 4 | Assignment 2 |
| 2 | Data Collection and Presentation | Week 5 | Assignment 2 |
| 3 | Probability | Week 6 &7 | Assignment 2 |
| **MODULE 3: INFERENTIAL STATISTICS** | | | |
| 1 | Sampling | Week 8 & 9 | Assignment 3 |
| 2 | Hypothesis and Significance Testing | Week 10 &11 | Assignment 3 |
| 3 | Correlation | Week 12 | Assignment 3 |
| 4 | Regression Analysis | Week 13 | Assignment 3 |
| | Examination | Week 14 & 15 | |
| | **Total** | **15 Weeks** | |

## HOW TO GET THE MOST FROM THIS COURSE?

In distance learning the study units replace the university lecturer. This is one of the great advantages of distance learning; you can read and work through specially designed study materials at your own pace and at a time and place that suit you best.

Think of it as reading the lecture instead of listening to a lecturer. In the same way that a lecturer might set you some reading to do, the study units tell you when to read your books or other material, and when to embark on discussion with your colleagues. Just as a lecturer might give you an in-class exercise, your study units provides exercises for you to do at appropriate points.

Each of the study units follows a common format. The first item is an introduction to the subject matter of the unit and how a particular unit is integrated with the other units and the course as a whole. Next is a set of learning objectives. These objectives let you know what you should be able to do by the time you have completed the unit.

You should use these objectives to guide your study. When you have finished the unit you must go back and check whether you have achieved the objectives. If you make a habit of doing this, you will significantly improve your chances of passing the course and getting the best grade.

The main body of the unit guides you through the required reading from other sources. This will usually be either from your set books or from a readings section. Some units require you to undertake practical overview of historical events. You will be directed when you need to embark on discussion and guided through the tasks you must do.

The purpose of the practical overview of some certain historical economic issues are in twofold. First, it will enhance your understanding of the material in the unit. Second, it will give you practical experience and skills to evaluate economic arguments, and understand the roles of history in guiding current economic policies and debates outside your studies. In any event, most of the critical thinking skills you will develop during studying are applicable in normal working practice, so it is important that you encounter them during your studies.

Self-assessments are spread throughout the units, and answers are given at the ends of the units. Working through these tests will help you to achieve the objectives of the unit and prepare you for the assignments and the examination. You should do each self-assessment exercises as you come to it in the study unit. Also, ensure to master some major historical dates and events during the course of studying the material.

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

1. Read this Course Guide thoroughly.

2. Organise a study schedule. Refer to the `Course overview' for more details. Note the time you are expected to spend on each unit and how the assignments relate to the units. Important information, e.g. details of your tutorials, and the date of the first day of the semester is available from study centre. You need to gather together all this information in one place, such as your dairy or a wall calendar. Whatever method you choose to use, you should decide on and write in your own dates for working breach unit.

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

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

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

6. Work through the unit. The content of the unit itself has been arranged to provide a sequence for you to follow. As you work through the unit you will be instructed to read sections from your set books or other articles. Use the unit to guide your reading.

7. Up-to-date course information will be continuously delivered to you at the study centre.

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

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

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

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

12. After completing the last unit, review the course and prepare yourself for the final examination. Check that you have achieved the unit objectives (listed at the beginning of each unit) and the course objectives (listed in this Course Guide).

## TUTORS AND TUTORIALS

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

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

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

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

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

## SUMMARY

The course, Social Science Statistics (FSS211), exposes you to basic statistics and the course guide gives you an overview of what to expect in the course. The course will therefore guide and teach you the statistical tools used by social scientists, researcher and policy makers in analysing issues upon which public policy decisions are based. The course will therefore be highly important to you as a researcher and on the successful completion of the course, you would have developed critical thinking skills with the material necessary for efficient and effective discussion on Social Science Statistics.

However, to gain a lot from the course please try to apply anything you learn in the course to term papers writing in other social science courses. We wish you success with the course and hope that you will find it interesting and useful.

**MAIN COURSE**

**CONTENTS** **PAGE**

# MODULE 1 INTRODUCTION, CONCEPTS AND METHODS OF STATISTICS



## UNIT 1 MEANING, TYPES, CONCEPTS AND IMPORTANCE OF STATISTICS

**CONTENTS**



## 1.0 INTRODUCTION

In today's global world where technology is advancing rapidly, numerical facts are continually sought after for effective and efficient management of scarce resources so as to promote growth and development be it individually as an individual or collectively as a society. However, obstacles may arise when these numerical facts are not understood or are interpreted wrongly. Therefore, it is very important to understand these figures properly. As a social scientist interested in the society and social relationships, the knowledge of these numerical facts, how to source for them, their interpretations, usage, and applications cannot be overemphasized and this is basically what statistics is all about.

In this unit, basic concepts of statistics are introduced and explained. These will lay the groundwork for the whole of the course material.

## 2.0 OBJECTIVES

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

- define Statistics
- discuss freely types of Statistics
- explain the relevance of Statistics to Social Scientists

## 3.0 MAIN CONTENT

### 3.1 Meaning of Statistics

Statistics is encountered in our everyday activity, the population figures of countries, the number of male and female students in a class or school, the age range of students in your class, and so on are all number based information we come across. However, this is just an aspect of statistics which deals with collation of quantitative or numeric data. When these everyday figures on a particular subject matter, objects or person are compiled, organised and put together, we have a comprehensive body of numerical information which is also collectively referred to as statistics. For example, we have Population statistics, health statistics, and housing statistics.

Let us consider the following information from the National Bureau of Statistics (2018).

1. In the year 2016, 41% of Nigeria's population was under age 15
2. Nationally, of the 742,488 births registered in 2016, 33.09 % of them were registered before age one.
3. Between 2013- 2015, out of the total 2912 recorded cases of people trafficked, persons within 6-15 age bracket were victims of trafficking in person more than any other age groups with 48%, followed by persons within 16-25 age bracket with 39%.

The figures contained in the information, that is 41%; 742, 488; 33.09%; 2912; 48% and 39% represent statistics, as they give us information or facts in figures on some demographic matters in Nigeria in the stated years. In this regards statistics can be defined as the numerical values or indicators computed through mathematical manipulation of the numerical data. This data can be presented in various form such as percentages, index numbers, averages, and medians. It can therefore be said to be the numerical measure that describes the characteristic of a sample.

Apart from being related to numeric information or data, Statistics can also be regarded as a discipline just like Economics, and someone who

is a professional in this field of study is called a Statistician just like an Economist is referred to a professional in the field of Economics. As a field of study, Statistics can be defined as that science or branch of learning which deals with the method of collection, analysis, presentation, as well as drawing conclusions from the numeric data (Obikeze, 1986).

As a social scientist, you are concerned with the issues in the society and as such, you are expected to make use of social science data like data on population. When the science of statistics is applied to these social science data, it is called social statistics. In Social science, the science of statistics or put simply, statistics is a tool and therefore not an end but a, means to an end. This means that it is not sought after for the direct benefit it provides but for what it is used for and in social science it is used for solving social issues in our society. In a later section, we will discuss the importance of statistics in social sciences.

**SELF-ASSESSMENT EXERCISE**

Differentiate between statistics as numeric values and statistics as a field of study.

## 3.2 Types of Statistics

In the field of statistic, there are two types or branches of statistics and these are *descriptive* and *inferential* statistics. According to the SAGE dictionary of statistics, descriptive statistics is a wide variety of techniques that allow us to describe the general characteristics of the data we collect. Note that just like the name implies, it basically describes our data. With descriptive statistics, we can summarise and describe our data without making a general statement or conclusion from the data. We describe these data using percentages, averages or mean (and other measurements of central tendency which will be learnt in unit 1 Module 2), and various forms of graphical representations like tables, charts and graphs. Then the question is, what then do we do with the described data? This is where the second type of statistics comes into play and that is the inferential statistics.

Inferential statistics, going by the definition from the SAGE dictionary, is that branch of statistics which deals with generalisation from samples to the population of values. Population here means the total collection of objects that the researcher is interested in studying while the sample is the subset of the population. The sample is a given portion of the population, selected scientifically, this selected portion taken from the population is called the subset of the entire population. With inferential

statistics, data is taken from samples to make generalisations about the population and this requires significant testing.

If for example as a social science researcher, you have been called upon to carry out a research on the opinions of Nigerian's on the fairness of the 2019 Presidential election. What are you expected to do? First of all, you should have it at the back of your mind that there is no way you can conduct a survey which will capture every single person in the country given the massive population we have in Nigeria, there is no chance that this is going to happen. With the time, money and other resources to be considered, your best bet is to scientifically determine a sample size, and then gather data from this sample.

In our example, let us assume that 50% of the people interviewed are of the opinion that the election was fair, then the descriptive statistics will report the answer as it is, while the inferential statistics will say that 50% of Nigerians believe or are of the opinion that the election was fair, meaning that the sample size selected has been used to make inference about the population. In using inferential statistics, care should be taken not to end up selecting sample that will not represent the entire population that is to say, the sample, must give every member of the population an equal chance of being selected. If the selection of sample size is not properly handled, wrong results will be gotten leading to erroneous conclusions.

Sometime, population and sample can be the same, that is to say the entire population is captured, like we have in censuses. This is usually expensive and it is carried out centrally by the government and it is done at given time intervals, usually ten years' intervals. However, the last time a census was conducted in Nigeria was in 2006, the one before that was in 1991, which is a period of fifteen years' interval. With census, everyone is captured because its aim basically is to gather data on all citizens of a country.

To throw more light on the two types of statistics, we can use everyday examples of what people say. Let us look at these three examples.

1) Kaychi's Cumulative Grade Point Average (CGPA) is 3.25.
2) 60% of Emmanuel's classmates are men.
3) The average age of the people in Emmanuel's class is 28 years.

We hear these types of statements all the time and they basically represent the summary of the total data presentation. Like in the case of the CGPA in example one, it describes Kaychi's average academic performance, so it is a descriptive statistic. Then in the third example, which is the average age of the people Emmanuel's class, this average is

gotten by summing up all the ages of Emmanuel's classmates and then proceed to divide this summed up value by the number of people in his class. This describes the average age of his classmates and basically gives us a picture of what the age of the members of the class is, how young or old they are. By this we have an idea of the age group of the people in the class. So the average/ mean was used to give this outlook and the job of descriptive statistics ends here. These averages given in these examples (examples 2 and 3) are just for the particular class and cannot be used to make conclusions for the whole school and as such it would be wrong to say the average age of the people in Emmanuel's school is 28 or 60% of Kaychi's school mates are men. When generalization is to be made, then we have gone beyond descriptive and are in the domain of inferential statistics.

**SELF-ASSESSMENT EXERCISE**

Differentiate between descriptive and inferential statistics

## 3.3 Basic Statistical Concepts

Like all other disciplines, statistics has its own terminologies and words peculiar to it. You must have encountered some of these terminologies in the first few sections of this work and will still come across, them and others as you proceed in this course. The basic terms are listed and briefly defined below for you to familiarise yourself with them. We will not elaborate on these words in this section because some of them, like population and sample have been discussed in previous sections and the others will be discussed in sections to come. The definitions of terms are listed below:

**Population:** In statistics, population refers to *all* of a particular type of individual or objects to be studied. This may be limited by geographical location or one or more other characteristics. Therefore, populations would include all NOUN students, all the students in Nigerian Universities, or all vegetarians in the country, etc.

**Sample:** This is a set of cases drawn or selected from a larger set or population of cases, usually with the aim of estimating characteristics of the larger set or population.

**Statistic:** A characteristic such as a mean, standard deviation or any other measure which is applied to a sample of data. Applied to a population exactly the same characteristics are described as parameters of the population.

**Parameter:** Is a numerical characteristic of a population or a model.

**Data:** With plural form as datum, it is the information or facts about something.

**Variable:** This is a characteristic that consists of two or more categories (such as occupation or nationality) or values (such as age or intelligence score). The opposite of a variable is a constant, which consists of a single value.

**SELF- ASSESSMENT EXERCISE**

In statistics, some basic terminologies are used. Mention any three you know.

## 3.4 Importance of Statistics to Social Scientists

Statistics is vital in our daily activity because it keeps us informed on what is happening around us. The knowledge of statistics enables us to make intelligent choices in whatever field we find ourselves be it in Health, education or Agriculture.

In Business, the knowledge of statistics will assist in making decisions on issues like where to site a business, what to produced and price to fix based on the data gotten from market analysis. In health sector, it will aid health practitioners to know the efficacy of drugs developed, to know how to trace and control epidemics, to know how to better inform the public and policy makers on health issues in order to promote well-being and development, and a whole range of other benefits.

As a social scientist, some of the importance of statistics are listed below.

- As a social scientist, the knowledge of statistics will aid you in evaluating the credibility and usefulness of any information you come across, and this will guide you in making the right decisions, be it as a consumer or a producer of goods and services.
- The knowledge of statistics will help you the learner become a more informed user of research tools and statistical analyses that affect many aspects of your life.
- An understanding of statistics will aid the learner in carrying out his or her own research.
- With the aid of statistics, raw data are summarised and made comprehensible. That is to say, the knowledge of statistics helps to give meaning to data thereby making it usable.

- With the knowledge of statistics, the social scientist is able to know how and under what conditions to make inferences from a small group to a larger group.
- The social scientist is able to apply the knowledge of statistics in handling social science data so as to understand the relationship between variables and the extent to which they are related. In social sciences this is important because with this knowledge, you will know how variables interact with one another. For example, the relationship between income and happiness; taste/preferences and prices of goods and services; conflict and religious beliefs, crime rate and unemployment, can be known through the use of statistical methods which will not only tell us the relationship but also tell us the extent of their relationship, the future expectations(forecasting) and a whole lot more.

**SELF- ASSESSMENT EXERCISE**

Mention two importance of statistics to you as a social scientist.

## 4.0 CONCLUSION

This unit has been able to expose you to what statistics is all about and why it is necessary to study it. Statistics has two definitions. It can be defined as numeric facts and also as a field of study. Whichever we define it, the crux of the matter is that it is a science which deals with numeric data and hence its importance in today's technologically driven world. Its importance in our everyday life and understanding it cannot be overemphasised as it is useful in diverse fields of learning. For a social scientist, the knowledge of statistics goes a long way in making us better decision makers in our society.

## 5.0 SUMMARY

To throw more light on what we have discussed so far about what statistics is, the unit noted that:

1. Statistics can be seen as both numeric data and a field of study.
2. There are two branches of statistics, which are the descriptive and inferential statistics.
3. While descriptive statistics is that branch of statistics which involves organising, displaying, and describing of data, inferential statistics is the branch of statistics that involves drawing of conclusions.
4. Some basic concepts in statistics which you will come across regularly as you study are population, parameters, variable, sample, data, and hypothesis.

5. As a social scientist, the knowledge of statistics will guide you in making sound decisions.

## 6.0 TUTOR-MARKED ASSIGNMENT

1. Explain the difference between descriptive and inferential statistics.
2. In a study carried out on how much families spend on school lunch at Darell Academy, 50 families with children in the school were randomly selected and from the survey, it was discovered that three families spent N70,000, N75,000 and N68,000 respectively. Identify the basic statistical concepts in the illustration, the first answer has been provided.

Population: All families with children at Darell Academy
Sample:
Parameter:
Statistics:
Variable:
Data:

## 7.0 REFERENCES/FURTHER READING

Anderson, D. R., Sweeney, D. J. & Williams T. A. (2008). *Statistics* for Business and Economics. (10th ed.). USA: Thomson Corporation.

Cramer D. & Howitt D. (2004). *The SAGE Dictionary of Statistics a Practical Resource for Students in the Social Sciences.* SAGE Publications London.

National Bureau of Statistics. (2018). *National Bureau of Statistics 2017 Demographic Statistics* Bulletin. Accessed: https://nigerianstat.gov.ng/download/775.

Obikeze D. S (1986). *Introductory Statistics for the Social Sciences.* Fourth Dimension Publishing Co. Ltd, Enugu.

Tokunaga H. T. (2016). *Fundamental Statistics for the Social and Behavioral Sciences.*SAGE Publications, California.

# UNIT 2 MEANING, TYPES AND FUNCTIONS OF STATISTICAL SYMBOLS

**CONTENTS**



## 1.0 INTRODUCTION

There are some basic statistical signs and symbols which form part of the language in statistics, these signs and symbol are often used for problem solving and are divided into three groups: alphabetical, Greek and mathematical symbols. This unit introduces you to some of these signs and symbols. You are expected to take note of them as you will come across them not only in this course but in all your statistical and mathematical related courses as well as some calculations, especially the scientific ones.

## 2.0 OBJECTIVES

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

- explain what statistical symbols mean
- explain the various types of statistical symbols
- state and analyse the basic functions of mathematical symbols.

## 3.0 MAIN CONTENT

### 3.1 Basic Statistical Symbols

Statistical symbols have their various meanings and what they stand for. Table 1.1 shows some common symbols used in field of Statistics. Their meanings will be explained in section 3.2.

**Table 1.1: Some Basic Statistical Symbols**

| Symbol | Symbol Name |
|---|---|
| CV | Coefficient of Variation |
| P(x) | Probability of x |
| M | Mu |
| Σ | Sigma (lower case) |
| ∓ | Minus-plus |
| Df | Degree of Freedom |
| Γ | Gamma |
| ( ) | Parentheses |
| … | Ellipsis |

***Source:*** *Author's Compilation, 2019*

**SELF- ASSESSMENT EXERCISE**

Give two basic statistical symbols you know.

## 3.2 Types and Meanings of Some Statistical Symbols

There are three different types of statistical symbols and signs. These symbols are classified based on their representation or origin. The three types of signs and symbols in statistics are the alphabetical statistical symbols, the Greek alphabetical symbols and the mathematical alphabetic symbols. These symbols are shown in tables 1.2, 1.3 and 1.4.

**1. Alphabetical Symbols**

These statistical symbols are alphabetical symbols with alphabets like a, b, Y, and X. They are usually found in mathematical and statistical operations. For example, in plotting a graph, X and Y are used to represent the horizontal and vertical lines of the graph. While the horizontal line is labelled X (x-axis), the vertical line is Y(Y-axis). In statistics, X and Y are used for variable representation. These letters can represent variables like age, income, educational attainment, employment status etc. Superscripts and subscript can be added to these letters to give it various meanings. For example, $X^2$ stands for X multiplied by X, or X squared, and $X1$, $X_2$, $X_3$ are generally used to represent particular observations.

Also in statistics, most often, N stands for the sum of all cases and Y stands for the dependent variable in a model. Table 1.2 shows some of these alphabets and their meaning.

**Table 1.2: Some Alphabetical Symbols used in Statistics**

| **Some Alphabetical Symbols used in Statistics** | | |
|---|---|---|
| **Symbols** | **Symbol Name** | **Meaning/ Definition** |
| n | | Sample Size |
| N | | Population size |
| $\bar{x}$ | X- bar | Arithmetic Mean |
| $S^2$ | s-squared | Sample Variance |
| S | $S^2$ | Standard deviation |
| SD | | Standard deviation |
| $H_o$ | H-naught | Null Hypothesis. The independent variable has no influence on the dependent variable |
| $H_i$ | H-one | Alternate Hypothesis. An alternate hypothesis is accepted when the null hypothesis must be rejected |
| p-Value | | The attained level of significance |
| R | | Sample Correlation |
| $R^2$ | R-Square | Multiple correlation coefficient |
| $r^2$ | r-square | Coefficient of determination |
| $t_c$ | T critical | The critical value for a confidence level c |
| X | | Independent variable in regression analyses |
| Y | | Dependent variable in regression analyses |
| $Z_c$ | z critical | The critical value for a confidence level c |

***Source: Author's Compilation, 2019***

2. **Greek Symbols**

It is common to find Greek symbols in statistical operations; these symbols are gotten from the Greek alphabets. One very common Greek symbol used in statistics is the Greek alphabet called Beta, written as β. In regression analysis, the beta value is a parameter which measures the relationship and how strongly each independent variable influences the dependent variable. Some of these Greek symbols are shown in table 1.3.

**Table 1.3: Some Greek Symbols Used in Statistics**

| **Some Greek Symbols Used in Statistics** | |
|---|---|
| Greek Symbol | Symbol Name |
| **α** | Alpha |
| β | Beta |
| | Epsilon |
| ρ | Rho |
| $X^2$ | Chi-square distribution |
| Σ | Sigma (Upper case) |
| λ | Lambda |

***Source: Author's Compilation, 2019***

### 3. Mathematical Symbols

Mathematical symbols are what we have been introduced to in our early learning years right from elementary school. They are symbols we come across on regular bases and as such we may have little or no difficulties identifying them. Table 1.4 shows some of these symbols which are commonly used in statistics.

**Table 1.4: Some Mathematical Symbols Used in Statistics**

| **Some Mathematical Symbols Used In Statistics** | | |
|---|---|---|
| **Symbols** | **Symbol Name** | **Meaning/ Definition** |
| + | Plus | Addition |
| _ | Minus | Subtraction |
| ÷ | Divide | Division |
| × | Multiply | Multiplication |
| = | Equals | Equality |
| ≠ | Inequality | Not equal to |
| > | Strict inequality | Greater than |
| < | Strict inequality | Less than |
| ≥ | Inequality | Greater than or equal to |
| ≤ | Inequality | Less than or equal to |
| % | Percentage | Percentage |
| … | Ellipsis | Continuation of pattern |
| * | Asterisk | Multiplication |
| | Union | Or |
| | Intersect | And |
| ^ | Caret | Exponent |
| → | Arrow | Causes/ Influences |

***Source: Author's Compilation, 2019***

**SELF- ASSESSMENT EXERCISE**

What are the three different ways you can consider statistical symbols?

## 3.3 Basic Functions of Mathematical Symbols

Statistical symbols are used for performing operations in statistics. These symbols have been grouped based on the types of operations they perform. Based on these operations, the symbols have been grouped under connectives and operators.

**Connectives:** the term connectives is derived from the English word and just like the name implies, its function is to connect, link or join whatever needs to be joined such as sentences, clauses etc. In statistics this is also applicable as some symbols are used to link or specify the relationship between two or more variables or entities. Some of these symbols are given below.

**Table 1.5: Basic Connectives in Statistics**

| Symbol | Symbol Name | Example |
|---|---|---|
| = | Equal | 4 = 3+1 |
| < | Less Than | 3 < 4. Means 3 is less than 4 |
| ≠ | Not Equal to | 4 ≠ 3. Means 4 is not equal to 3 |
| → | Arrow | Y → X. Means Y influences X |
| > | Greater Than | 4 > 3. Means 4 is greater than 3 |

***Source: Obikeze (1986)***

**Operator:** A symbol which gives the actual instructions on what to do. These symbols tell us what is required of us when numerical figures are presented. See Table 1.6 below.

**Table 1.6: Basic Operator in Statistics**

| **Symbol** | **Symbol Name** | **Example** |
|---|---|---|
| Σ | Sigma | $\sum_{n=1}^{n} = 1 + 2 + 3 + 4 = 10$ |
| + | Plus | x +3 = 5 |
| – | Minus | 2-3 = -1 |
| ÷ | Divide | 4÷2 = 2 |
| × | Multiply | 2x3 = 6 |

Source: Obikeze (1986)

**SELF- ASSESSMENT EXERCISE**

Solve $\sum_{n=1}^{4} n^2$

## 4.0 CONCLUSION

In this unit, statistical symbols were discussed. Different symbols were introduced and explanations were given. Also, you were taught that these signs and symbols are grouped into three based on their types. These groups are the alphabetic, Greek and mathematical symbols. In addition to these, the symbols were also categorised into the function they perform in statistical operations. Two categories were given and they are the connectives and operators. All the different categories of symbols were tabulated to give clear and easy information to aid your understanding of the topic. The target is to aid you in identifying and

understanding better the statistical signs and symbols you will come across in other sections of the course materials as you proceed.

## 5.0 SUMMARY

This unit covered topics on the various statistical signs and symbols. The under listed key points were discussed extensively in this unit.

1. Signs and symbols form part of the language in statistics.
2. These signs and symbols are grouped into three. The three groups are the Alphabetic, Greek and Mathematical symbols.
3. The Greek symbols are gotten from the Greek alphabets and examples are the Greek alphabets α- Alpha**,** β- Beta, and Σ- Sigma.
4. Based on their statistical operations, these symbols can also be categorized into connectives and operators.
5. Examples of connectives are the equal and greater than symbols, while example of the operatives are the addition and subtraction signs.

## 6.0 TUTOR-MARKED ASSIGNMENT

1. List and give examples of the two types of statistical symbols gotten from the Greek alphabets.
3. Statistical symbols can be grouped based on the statistical operations they perform. Discuss.

## 7.0 REFERENCES/FURTHER READING

Anderson, D. R., Sweeney, D. J. & Williams T. A. (2008). *Statistics for Business and Economics,* Tenth Edition. USA: Thomson Corporation.

Cramer D. & Howitt D. (2004). *The SAGE Dictionary of Statistics a Practical Resource for Students in the Social Sciences.* London*:* SAGE Publications.

Everitt, B. S. &Skrondal A. (2010). *The Cambridge Dictionary of Statistics (*4th ed.). London: Cambridge University Press.

Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimmer, H., (2014). *Introduction to Statistics*. Rice University, Houston, TX.Accessed:http://onlinestatbook.com/Online_Statistics_Education.pdf

Levine, D. M. & Stephan, D. F. (2005). *Even You Can Learn Statistics. A Guide for Everyone Who Has Ever Been Afraid of Statistics*.

Pearson Education, Inc. Publishing as Pearson Prentice Hall Upper Saddle River, NJ. USA Accessed: file:///C:/Users/1/Downloads/[David_M._Levine,_David_F._Stephan]_Even_you_ can_l(z-lib.org)%20(1).pdf.

National Bureau of Statistics (2018). *National Bureau of Statistics 2017 Demographic Statistics:* Bulletin. https://nigerianstat.gov.ng/download/775.

Obikeze D. S (1986). *Introductory Statistics for the Social Sciences.* Enugu: Fourth Dimension Publishing Co. Ltd.

Shafer, D. S. & Zhang, Z. (2012). *Introductory Statistics*. Flat World Knowledge. Washington, DC, USA.

Tokunaga H. T. (2016). *Fundamental Statistics for the Social and Behavioral Sciences*. California: SAGE Publications.

Watkins, J. C. (2016).*An Introduction to the Science of Statistics*: From Theory to Implementation (Preliminary Edition). Accessed: https://www.math.arizona.edu/~jwatkins/statbook.pdf.

# UNIT 3 VARIABLES AND THEIR MEASUREMENTS

## CONTENTS



## 1.0 INTRODUCTION

This study unit explains the meaning of a variable. From the definition and explanation of variables, the student will know the importance of variable in research analysis. The students are introduced to these in order to have the full grasp of what variables are as it will be discussed and employed extensively in the entire course work.

## 2.0 OBJECTIVES

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

- define a variable
- discuss types and classes of variables
- explain the different ways to measure a variable.

## 3.0 MAIN CONTENT

### 3.1 Variables and their Measurements

In statistics, a variable is quantity or variate that can assume different values and can be measured or counted. The world today is characterised by and composed of objects that are directly observable through senses and ideas referred to as concepts. These objects and concepts have different features, attributes or properties used to identify them; for example, cars differ in size, colour and cost while bread differs in taste, size etc. These varying attributes is what is known as variable. Variables therefore are properties or characteristics of some event, object, or person that can take on different values or amounts as

opposed to constants. A house is not a variable; the shape and cost of the house is the variable. When conducting research, researchers often manipulate variables. Note that variables provide the raw materials for statistical operations and all variables taken together form the data of an analysis. The specific value of a variable is referred to as an observation, a score, a case or an item. For example, in a research like "the social status of 2000 B.Sc Development Studies students in National Open University of Nigeria", social status is the variable and the status of each student concerned is the observation or case.

**SELF-ASSESSMENT EXERCISE**

Explain the term variable.

## 3.2 Classification of Variables

Variables are grouped into several categories and they include;

a. Quantitative and Qualitative Variables
b. Independent and Dependent Variables

### 3.2.1 Quantitative and Qualitative Variables

**Quantitative variables**
Quantitative variables are variables which take on values that vary in terms of magnitude. Their values result from counting or measuring. This implies that they exist in greater or lesser amount and the observations have measurable magnitude. Quantitative variables are measured and expressed numerically. They have numeric meaning, and can be used in calculations. It is thus referred to as numerical variables. Examples of quantitative variables include; height, weight, distance, age, annual income, amount of rainfall, scores in a test, interest rate, etc. Quantitative variables are further classified as *discrete* or *continuous*.

**Discrete Variables**
Discrete variables are quantitative variables whose values are expressed as whole numbers or rounded numbers. They can only take on a finite number of values. This means that the numbers consist of only integers as they do not exist in fractional part in real life. Discrete variable are therefore countable variables. When we count things, we use whole numbers like 0, 1, 2, and 3. For example, you can count the change in your pocket. You can count the money in your bank account. Furthermore, we can count how many eggs a chicken lays. We know that each day a hen may or may not lay an egg, but there are two things that can never happen. There can never be a negative number of eggs, and there can never be a fraction or a portion of an egg. It is always a

whole number hence; it is a discrete variable. Examples include family size, goals scored in a game, number of coin flips, number of books published, number of rooms in a hotel etc.

**Continuous Variables**

Continuous Variables are numeric variables whose values assume theoretically an infinite amount of gradation between any two values. Continuous variable thus takes on infinitely, uncountable values. In continuous variable, just as the name implies, it takes forever to count. In fact, you would get to "forever" and never finish counting them. Example of continuous variable is age. In the case of age, there is no finite age. We thus have ages in forms like 2 years, 10 months, 5 hours, 4 seconds, 4 milliseconds, 8 nanoseconds, 99 picoseconds. Other examples include

**Qualitative Variables**

Qualitative variables are variables that take on values that vary in kind rather than magnitude. Qualitative variables that are not measurement variables thus their values do not result from measuring or counting. In other words, they differ in types or attribute rather than in quantity. Qualitative variable includes all observable qualities or characteristics of a group or population. Examples include hair color, religion, political party, profession. Qualitative variables are grouped into two types: nominal and ordinal(lacking a criterion of order) and ordinal (they have a criterion of order).

**Nominal qualitative variables**

Nominal qualitative variables are those that lack or do not admit a criterion of order and do not have an assigned numerical value. An example of such variables may be marital status (married, single, divorced, widowed).

**Ordinary qualitative variables**

Ordinary qualitative variables are known as semi-quantitative variables. Although they allude to attributes or qualities that lack a numerical value, they are classified within a scale of value. An example of this type of variables can be the result of a sport competition (first, second or third place).

### 3.2.2 Independent and Dependent Variables

Here, variables are classified based on the nature of relationship existing among them. This classification of variables is research oriented. It is important for you to know that the classification is based on the function of the variable in the research and thus, no variable is totally dependent or independent. A variable may be dependent in a circumstance and become independent in another circumstance.

**Independent Variables**
An independent variable is a variable that influences the behavior of another variable. It represents inputs or causes. It is used to predict the value of another variable. It is thus a variable which we assign value and whose variation does not depend on that of another variable. It is denoted with the symbol X. it is the variable that a researcher can manipulate in a study. For example, income is an independent variable because it causes and influences another variable called consumption. Also called regressor(s), controlled variable(s), explanatory variables(s) or predictor variable(s).

**Dependent variables**
A dependent variable is what is being measured in a research. The dependent variable is sometimes called the outcome variable and it is the variable being predicted. Thus, the value of the dependent variable depends and is determined by the independent variables and other factors. For example, a test score could be a dependent variable since it depends on several factors such as how much you studied, how much sleep you got the night before you took the test, etc. Also, when one is doing chores to earn some allowance, the dependent variable is the amount of money you earn because the amount of money you earn depends on how many chores you do. It is denoted with the symbol Y.

In the case of the independent and dependent variables, the functional relationship is stated thus:
$Y = f(X)$, i.e. Y depends on X
Thus, X is the independent variable and Y is a dependent variable. Note that in plotting a graph, the dependent variable is placed on the vertical (y-axis) while the independent variable is placed on the horizontal (x-axis).

There are two classes of variable we will discuss below. These variables are quite similar to dependent variables and are sometimes used in statistical analysis. They are endogenous and exogenous variables.

**Endogenous variables**

Endogenous variables are used in regression analysis. They are similar to but not exactly the same as dependent variables. Endogenous variables have values that are determined by other variables in the system. Therefore, a variable is said to be endogenous within the causal model if its value is determined or influenced by one or more of the independent variables excluding itself. However, in real life purely endogenous variables are rare.

**Exogenous Variables**

An exogenous variable is a variable that is not affected by other variables in the system. It is the opposite of endogenous, and therefore, it is a variable that is not influenced by any other variables in the model of interest. In other words, the value is not determined in the system being studied. The value of an exogenous variable is fixed in model that is, it is not determined within the model. Exogenous variables are not explained by the model rather; they influence the endogenous variables in the model.

**SELF-ASSESSMENT EXERCISE**

i. List and discuss the classification of variables.
ii. List the groups of qualitative variables.

## 3.3 Measurement of Variables

Instruments for measuring variables are referred to as measurement scales or levels of measurement and since variables are of different kinds, there are several ways of measuring them. Of all types of measurement, four scales or levels of measurement are popularly used. These types are discussed below

i. **Nominal Scale**

This is the lowest and simplest scale of measurement. Nominal scales simply sort the objects into categories and assign numbers as labels to identify objects or classes of objects. The assigned numbers have no meaning except as identifiers. The scale separates variables into mutually exclusive categories and attaches a label or name to each group. For example, the use of ID codes A, N and P to represent aggressive, normal, and passive drivers, is a nominal scale variable. Note that the order has no meaning here, and the differences between identifiers are meaningless. In practice it is often useful to assign numbers instead of letters to represent nominal scale variables. Example of nominal scale variables are sex, occupation, nationality etc.

ii. **Ordinal Scale**

Ordinal scales like the nominal scales sort variables into mutually exclusive categories, assigning numbers to objects and in addition, rank them. The ordinal scale thus categorises, labels and/or ranks the variables. In the ordinal scale, note that the classes must be put into an order such that each case in one class is considered greater than (or less than) every case in another class. Cases in the same class are considered to be equivalent. Although order does matter in these variables, the difference between responses is not consistent across the scale or across individuals who respond to the question. Examples of ordinal scale include movie ratings, political affiliation, military rank, etc. These scales are generally used in research to gather and evaluate relative feedback. For example, a scale question such as: How satisfied are you with our services? With expected feedback based on customer satisfaction categorised into

- Very Unsatisfied – 1
- Unsatisfied – 2
- Neutral – 3
- Satisfied – 4
- Very Satisfied – 5

iii. **Interval scale**

Interval scale contains all the properties of the ordinal scale. However, in an interval scale, an addition of the distance between any pair of adjacent categories is known. That is, it offers a calculation of the difference between variables. The interval scale therefore provides the information on how much one category is more or less than the other. In addition to which, the main characteristic of this scale is the equidistant difference between objects. For example, the population density in four areas in an urban town is given as 1500, 700 and 300. Here, the distance between the values of any two of the areas is known and we can say not only that the second area has a higher density than the third but also that its density exceeds the other by 400. Example of interval variables are age expressed in years, income, temperature etc.

iv**. Ratio scale**

This is the highest of the scales discussed so far. Ratio scales have the properties of all the three scales discussed above. Inaddition, it has an absolute “zero” point. For example, traffic density (measured in vehicles per kilometer) represents a ratio scale. The density of a link is defined as zero when there are no vehicles in a link. Other ratio scale variables include number of vehicles in a queue, height of a person, distance traveled, accident rate, etc.

**SELF -ASSESSMENT EXERCISE**

List the different ways that variables could be measured.

## 4.0 CONCLUSION

This unit has been able to expose you to what variables are all about and why it is necessary to know the classes and measurement of these variables. Variable therefore are properties or characteristics of some event, object, or person that can take on different values or amounts as opposed to constants. Their importance is that they are useful in diverse fields of social sciences and they drive the research process.

## 5.0 SUMMARY

To throw more light on what we have discussed so far about what variable is, the unit notes that:

1. A variable is a quantity that can assume different values and can be measured or counted.
2. variables provide the raw materials for statistical operations and all variables taken together form the data of an analysis.
3. The specific value of a variable is referred to as an observation
4. Variables are broadly grouped into quantitative and qualitative variables and independent and dependent variables
5. Instruments for measuring variables are referred to as measurement scales or levels of measurement
6. There are four well known scales or level of measurement

## 6.0 TUTOR-MARKED ASSIGNMENT

1. Explain what you understood by the term variable.
2. Explain what is meant by endogenous and exogenous variables.
3. Discuss the different groups of qualitative variables.

## 7.0 REFERENCES/FURTHER READING

Carlson, R. (2006). *A Concrete Introduction to Real Analysis*. CRC Press.

Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms.* Oxford University Press.

Everitt, B. S. (2002). The Cambridge Dictionary of Statistics (2nd ed.). Cambridge University Press. ISBN 0-521-81099-X.

Kozak, A., Kozak, R., Watts, S., &Staudhammer, C. (2008). *Introductory Probability and Statistics*. Vancouver: Cabi International.

Obikeze D. S (1986). *Introductory Statistics for the Social Sciences.* Enugu: Fourth Dimension Publishing Co. Ltd.

Okoro, E. (2002). *Quantitative Techniques in Urban Analysis*. Ibadan: Kraft Books Ltd.

# MODULE 2 DESCRIPTIVE STATISTICS AND PROBABILITY



## UNIT 1 MEASURES OF CENTRAL TENDENCY AND DISPERSION

**CONTENTS**



## 1.0 INTRODUCTION

Recall that in Module 1, we discussed the two branches of statistics which are the descriptive and inferential statistics. In this unit, we will focus on the major types of descriptive statistic. Our emphasis shall be on the two key types of descriptive statistics which are the measures of central tendency and the measures of dispersion. In Statistics, these measures are needed to give meaning to data, without them, the data collected will provide no meaningful information.

The measures of central tendency are used to identify the number which represents the center or the middle of a set of data. Meanwhile, the measures of dispersion used to find the spread how close or far the other

values are from the center. Under the measures of central tendency, the three most common measures are the mean the median and the mode. These will be discussed. So also, we shall discuss the, range, mean deviation, variance and standard deviation under the measures of dispersion.

## 2.0 OBJECTIVES

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

- explain the various measures of central tendency and dispersion
- calculate the mean, mode and median of a data set.
- Calculate the range, the standard deviation and variance of a data set.
- identify Outliers in a data set.

## 3.0 MAIN CONTENT

### 3.1 Measures of Central Tendency

When data are gathered, they are called raw data because they have no meaning and therefore make no sense to anyone. To be able to use data for the purpose they are gathered, the data have to be summarised into a single index or value which will truly represent the entire data set. One of the ways this is done is by using the measures of central tendency. The measures of central tendency are those measures or index which describes the central value in a data set. Since the central value is in the middle, all the other values in the distribution are arranged around it.

The most common measures of central tendency are the mean, the mode and the median. These three indices are the same when the distribution is unimodal and symmetrical. These indices are particularly interested in locating the average or center of the data set. The reason why the central location of the data set is important in statistics is because a lot of variables studied by researchers have majority of data clustered in the middle of the distribution. Therefore, it makes great sense to start an analysis by identifying the center of the distribution in a bid to describe or summarise the data. The mean, mode and median are discussed below.

#### 3.1.1 The Mean

The mean which is most often regarded as average, is the most common measure of central tendency. It answers the question: what is the average value of the variable in this set of data? Thus, it provides a measure of central location for the data. The mean is sometimes referred to as the

arithmetic mean in order to differentiate it from other types of mean like the geometric and harmonic mean, however because it is so commonly used it is simply referred to as the mean. To calculate the mean of a distribution, the sum of scores in the distribution or data set is divided by the number of scores. The formula for calculating the mean denoted by $\bar{x}$ (called X - bar like you were taught in section 3.2 in unit one) is $\frac{\sum x_i}{n}$

Where: n = number of observations in a sample

Σ = summation (see table 1.3 in section 3.2)

$x$ = value of the variable in the distribution, with $x_1$, $x_2$ representing the values of the first and second observations respectivelywhile $x_i$ represents the value of variable of the $i^{th}$ observation.

From the formula therefore the summation of the variables will give:

$\sum x_i = x_1 + x_2 + x_3 + x_4, \ldots + x_n$

To illustrate the computation of a sample mean, let us consider the following question.

**Question 1:** Find the mean of 3, 5, 7, 6, 8, and 4.

**Answer:**

The sum of the observations, $\sum x_i = 3 + 5 + 7 + 6 + 8 + 4 = 33$

The mean $\frac{\sum x_i}{n} = \frac{33}{6} = 5.5$

**Question 2:** The following numbers represent the ages of children at a party **(1; 2; 4; 3; 5; 6; 7; 8; 6; 5).** Calculate their mean of their ages.

**Answer:** In the question, number of observations (n) = 10

The sum of the observations, $\sum x_i = 1 + 2 + 4 + 3 + 5 + 6 + 7 + 8 + 6 + 5 = 47$

The mean $\frac{\sum x_i}{n} = \frac{47}{10}$

= 4. 7

The mean age of the children at the party is 4.7 years.

From Example 2, you may have noticed that two figures (5 and 6) appeared twice. Despite this double representation, the figures were all summed up just like we did in the first example. The first case represents the ungrouped data set, while the second case has a simple case of weighting, that is, the frequency. This can be seen with 5 with frequency of 2 and 6, with frequency of 2 as well. When some values are more in number than others, then the weighted mean is calculated. This is done by multiplying each weight (w) by its matching value (x) and summed up. The result of the computation is then divided by the sum of the weights.

Let us look at a complex case of weighted mean. Consider the following dataset:
2, 2, 2, 3, 3, 4, 4,1. To calculate the mean, we need to assign weights to the values as we can see that some numbers appear more than once. The best way to do this is to tabulate the data set.

**Table 1.1: Computing weighted Mean of a Data set.**

| S/N | Number (X) | Weight (W) | WX |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 2 | 3 | 6 |
| 3 | 3 | 2 | 6 |
| 4 | 4 | 2 | 8 |
| **Total** | | ΣW = **7** | ΣWX = **21** |

The weighted mean, $\frac{\Sigma WX}{\Sigma W} = \frac{21}{7} = 3$.
When data are presented in groups or placed in intervals, for example 20-24; 25-29 etc., the easiest method to use in calculating the mean is to represent each interval by its midpoint also called class mark, and then use this midpoint as the individual item or variable. In the case of the interval 20-24, the midpoint is 22.

The midpoints are then multiplied by their individual frequencies, then summed up and divided by the sum of their various frequencies.
The formula for the sample mean of a grouped data set is therefore given as:

$$\bar{x} = \frac{\sum mf}{\sum f}$$

Where:
$m$ = Interval midpoint
$f$ = Interval frequency

Note that the formula $\frac{\sum x_i}{n}$ as stated previously, represents the sample mean. As for the population mean, the computation remains the same, however, the formula has slightly different notation to accommodate the population factor. Thus, the formula for the population mean is given as:

$\mu = \frac{\sum x_i}{N}$ Where the μ = Population mean

N = number of observations in the population.

### 3.1.2 The Median

Just like the Mean, the Median is another type of measure of central tendency. It is defined as the value of a variable that splits a distribution of scores in half, with the same number of scores above the median as below it (Tokunaga, 2016). The median tells us what value of the variable is located in the middle or center of the data set when they are

written or arranged orderly in ascending order with the least value coming first. In calculating the median, if the number of observations is even, then the two values closest to the center are averaged. However, if the number of observations is odd then the number in the middle is taken as the median.

In order to understand this section, let us consider some examples.

**Example 1**: Compute the median in this data set. 9, 12, 7, 16, 14, 18 and 13.

**Solution**
First: arrange the data in an ascending order: 7, 9, 12, 13, 14, 16, and 18.

From the distribution above, it can be seen that the middle number is 13. Therefore, the median here is 13. This is easy to compute because the data set has odd number of observation which is 7. Let us consider a data set with even number of observation.

**Example 2:** A student scored 9, 7, 10, 8, 7, and 9 in 6 courses. What is the median score of the student?

The first step is to arrange the data set in an ascending order. This gives the following result: 7, 7, 8, 9, 9, and 10.

We can see that this set of data has even distribution, with the middle number lying between 8 and 9. To get the exact value, we are required to add the two values and then divide it by 2 to give the average/midpoint. This will give us a value of 8.5, therefore the median is 8.5.

### 3.1.3 The Mode

The mode is another measure of central tendency, just like the median and the mean. This is sometimes called the modal score. According to Everitt, and Skrondal (2010), the mode is the most frequently occurring value in a set of observations. As such, there can be a single mode when there is only one value occurring the most in the observation; no mode, when all values are equally represented in the observation and finally, there may also be more than one mode when there are two or more values occurring more frequently in the observation (see table 1.2).

The mode or modal score in a given distribution is usually identified visually, by looking at the frequency distribution table, or other visual representations like the use of the bar chart, graph or histogram. The

modal score will always be the one with the tallest bar in the chart (Tokunaga, 2016), or highest point on a graph etc. Perhaps, a solved example will make this clearer.

**Table 1.2: Examples of Modal Scores and what they are called**

| S/n | Data Set | Modal Score(s) | No. of Mode | Name Given |
|---|---|---|---|---|
| **Example:** 1 | 1,2,3,4,5,6,7 and 8 | 0 | 0 | - |
| **Example:** 2 | 1, 2, , 3,4,4,5,6,7, and 8 | 4 | 1 | Unimodal Distribution |
| **Example:** 3 | 1,2, 3, 3,4,4,5,6,7, and 8 | 3 and 4 | 2 | Bimodal or Multimodal Distribution |
| **Example:** 4 | 1,2,3,3,,4,4,5,6,6,7, and 8 | 3,4, and 6 | 3 | Trimodal or Multimodal Distribution |

**Example 1:** The following data represent the hourly wage (in Naira) offered to 10 caterers at ITSE Lodge, Jabi- Abuja. 5000, 2000, 3000, 5000, 5300, 5000, 3500, 4000, 3200, and 5300. What is the modal wage of the workers?

**Solution:** Tabulate the values in order to clearly show the frequency distribution.

From table 1.3, it can be clearly seen that the wage with the highest frequency is ₦5000.

**Table 1.3: Hourly Wages and Frequency Distribution of Caterers at ITSE Lodge, Jabi-Abuja.**

| Hourly Wage | Frequency |
|---|---|
| 2000 | 1 |
| 3000 | 1 |
| 3200 | 1 |
| 3500 | 1 |
| 4000 | 1 |
| 5000 | 3 |
| 5300 | 2 |
| Total | **10** |

### 3.1.4 Strengths and Weaknesses of the Mean, Median, and Mode

Having studied the three most common measures of central tendency, table 1.4 presents the weaknesses and strengths of these measures.

**Table 1.4: Strengths and Weaknesses of the Mean, Median, and Mode**

| Measure of Central Tendency | Strength | Weakness |
|---|---|---|
| **Mean** | Based on all of the scores in a set of data.<br><br>Can be used in statistical analyses to test hypotheses. | Affected by the extreme values in the distribution.<br><br>It cannot be calculated for an open-ended or truncated distribution.<br><br>May not describe bimodal distributions |
| **Median** | Not affected by extreme values. | Not based on all of the scores in a set of data.<br><br>Cannot be used in statistical analyses to test hypotheses.<br><br>May not describe bimodal Distributions. |
| **Mode** | It is not affected by the extreme values in the distribution.<br><br>Can be calculated for an open-ended or truncated distribution.<br><br>Quick and easy to Identify.<br><br>Can be used with categorical variables.<br><br>Describes bimodal or multimodal distributions | Not based on all of the scores in a set of data.<br><br>Cannot be used in statistical analyses to test hypotheses. |

***Source: Tokunaga, (2016) andObikeze (1986).***

**SELF-ASSESSMENT EXERCISE**

i. Find the mean, median, and mode of the data set: 2,4,3,9,6,8,8,4,4
ii. Differentiate between the mode, median and mean.

## 3.2 Measure of Dispersion

The measures of central tendency do not tell us anything about how scattered the observations are within the distribution. This leads to questions such as: how normal are your data? Does the sample represent the population well? Does the data have a regular pattern? Are the data precise or vague?

These questions are answered using the measures of dispersion**.** Measures of dispersion explain the degree to which numerical values tend to spread about its arithmetic mean value. The most common measures of dispersion include the range, mean deviation, variance, standard deviation and the interquartile range.

### 3.2.1 Range

The range specifies the distance between the highest and the lowest value in a given distributing. It is the difference between the highest value and the lowest value in a data set. Symbolically, the range is summarised as :

Range = H – L
Where H = Highest value
L = Lowest value

**Example**: What is the range in {3, 6, 11, 14, 3, 8},
Solution:
Range = H – L
H = 14
L = 3
So the range is = 14 − 3 = 11

**Advantages of using the range**
It is very easy to calculate and understand

**Disadvantages of Range**
i. It is easily affected by fluctuation of sampling. This implies that a variation in extreme value affects the range
ii. Since it does not consider all observations in the data, it may not be a good representative of such data.

iii. It is difficult to use in open ended distributions. If for any reason any of the extreme values is missing, the range becomes powerless

**SELF-ASSESSMENT EXERCISE**

i. Why is measure of dispersion essential in statistics?
ii. What is the range of the data set {7, 8, 7, 7, 4, 9, 11, 14, 15}?

## 3.2.2 Mean Deviation

The mean deviation is the average of the absolute deviation of a variable from its mean. The difference between a given case and the mean of the distribution is referred to as a deviate or deviation score. The mean deviation of a set of numbers $x_1, x_2, x_3 \ldots, x_n$ is symbolically denotes by

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

Where $\bar{x}$ is the arithmetic mean of the observation
The vertical lines in the equation | | are absolute value of the deviation $x_i$ from $\bar{x}$. This means that the mean deviation ignores the signs of any deviation, such that $|4 - 1| = |1 - 4| = 3$.
Note that if the observation $x_1, x_2, x_3 \ldots, x_n$ occur with frequency $f_1, f_2, f_3 \ldots, f_n$, then the mean deviation formula becomes

$$MD = \sum_{i=1}^{n} \frac{f_1|x_i - \bar{x}|}{\sum f_1}$$

**Example 1**: find the mean deviation of the following 4, 5, 2, 7, 9, 3
**Solution**

$$\bar{x} = \frac{4 + 5 + 2 + 7 + 9 + 3}{6}$$

$$\bar{x} = 5$$

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

$$= \frac{|4 - 5| + |5 - 5| + |2 - 5| + |7 - 5| + |9 - 5| + |3 - 5|}{6}$$

=

$$\frac{1 + 0 + 3 + 2 + 4 + 2}{6}$$

=

$$\frac{12}{6}$$

Mean Deviation (MD) = 2

**Example 2.** Calculate the mean deviation for the data set in the table below

| | 7 | 13 | 15 | 19 | 21 | 23 |
|---|---|---|---|---|---|---|
| Frequency | 4 | 4 | 3 | 2 | 4 | 6 |

Solution
Here,

$$\bar{x} = \frac{\sum f_1 x_1}{\sum f}$$

| $x_i$ | Frequency | $f_1 x_1$ | $\lvert x_i - \bar{x} \rvert$ From the formula, $\bar{x} = 1$ 9.3 6.3 | $f_1 \lvert x_i - \bar{x} \rvert$ |
|---|---|---|---|---|
| | 4 | 28 | 9.3 | 37.2 |
| 13 | 4 | 52 | 3.3 | 13.2 |
| 15 | 3 | 45 | 1.3 | 3.9 |
| 19 | 2 | 28 | 2.7 | 5.4 |
| 21 | 4 | 84 | 4.7 | 18.8 |
| 23 | 6 | 138 | 6.7 | 40.2 |
| **Total** | 23 | 375 | | 118.7 |

Since the question has frequency distribution for the data, then the formula below is applied in getting the mean deviation

$$\text{MD} = \sum_{i=1}^{n} \frac{f_1 \lvert x_i - \bar{x} \rvert}{\sum f_1}$$

$$\text{MD} = \frac{118.7}{23}$$

MD = 5.16

**Advantages of Mean Deviation**

i. All data points are put into consideration
ii. Since all data are considered, it is not affected by extreme value
iii. It is easy to calculate and understand

**Disadvantages of Mean Deviation**

i. It may be biased since it ignores the algebraic signs of deviation

**SELF-ASSESSMENT EXERCISE**

i. Calculate the mean deviation of the following 4, 5, 2, 7, 9, 3, 5, 7, 8, 9, 3, 11

## 3.2.3 Variance

The variance is the sum of mean deviation of the squares of the deviations measured from the mean. Essentially, the variance is a more precise measure of how precise your data is. It is represented by $s^2$ for a sample and $\sigma^2$ for a population. It is given as

$$S^2(\sigma^2) = \frac{\sum x^2}{N} = \frac{\sum (X - \bar{X})^2}{N}$$

**Note**: When the population or sample size is less (less than 30), the formula for variance is given as

$$\sigma^2 = \frac{\sum (x_1 - \bar{x})^2}{n - 1}$$

The following are the methods to follow when estimating the variance

i. Find the arithmetic mean of the observation
ii. Compute the deviation of data point from the arithmetic mean
iii. Square the deviation score, sum up the squared score and divide by the number of data point (N)

**Example:** Given the data set, calculate the variance 4, 5, 2, 7, 9, 3.

Solution

$$\sigma^2 = \sum \frac{(x_1 - \bar{x})^2}{n - 1}$$

$$\bar{x} = \frac{4 + 5 + 2 + 7 + 9 + 3}{6}$$

$$\bar{x} = 5$$

$$\sigma^2 = \frac{(4 - 5)^2 + (5 - 5)^2 + (2 - 5)^2 + (7 - 5)^2 + (9 - 5)^2 + (3 - 5)^2}{6 - 1}$$

$$\sigma^2 = \frac{(-1)^2 + (0)^2 + (-3)^2 + (2)^2 + (4)^2 + (-2)^2}{5}$$

$$\sigma^2 = \frac{34}{5}$$

$$\sigma^2 = 6.8$$

Note that if the observation $x_1, x_2, x_3 \ldots, x_n$ occur with frequency $f_1, f_2, f_3 \ldots, f_n$, depending on the sample size, variance becomes

$$S^2 = \frac{f_1(x_1 - \bar{x})^2}{n}$$

or

$$S^2 = \frac{f_1(x_1 - \bar{x})^2}{n-1}$$

**SELF-ASSESSMENT EXERCISE**

i. Calculate the variance of the dataset; 5, 4, 9, 3, 5, 2, 1, 4, 7.

### 3.2.4 Standard Deviation

This is perhaps one of the most widely used measure of dispersion. It is simply the square root of the variance and represented symbolically as *S* or σ. A high standard deviation means that the data set vary a lot, but a low standard deviation means that the data do not vary very much. the smaller the standard deviation, the better. Algebraically, it is given as:

$$\sqrt{\frac{\sum(x_1 - \bar{x})^2}{n}}$$

For sample standard deviation

or

$$\sqrt{\frac{\sum(x_1 - u_i)^2}{n}}$$

For population standard deviation

**Advantages of Standard Deviation**

i. Standard deviation is based on all the items in the series. So, it provides good representation of the data.
ii. Standard deviation can be used for mathematical operations and algebraic treatments. It is also applicable in statistical analysis.
iii. Unlike other measures, standard deviation is least affected by the sampling fluctuations

**Disadvantages of Standard Deviation**

i. Standard deviation is complex to compute and difficult to understand as compared to other measures of dispersion.
ii. Standard deviation is highly affected by the extreme values in the series.

iii. Standard deviation cannot be obtained for open end class frequency distribution

**SELF-ASSESSMENT EXERCISE**

Calculate the standard deviation of the data set 5, 4, 9, 3, 5, 2, 1, 4, 7.

## 3.3 Outliers

We studied the measures of central tendency in sections 3.1 and 3.2 of this unit. We were learnt that one of the weaknesses of the mean is that it is affected by the extreme values in the distribution. The mean therefore as a result of this weakness, does not accurately represent skewed distributions. These extreme values are what is called outliers. An outlier is defined as an observation that appears to deviate markedly from the other members of the sample in which it occurs (Everitt, and Skrondal 2010). Outliers present cases which are extremely unusual from the rest of the observations and as such may influence the interpretation of the data. Consequently, the result from this set of data may be misleading. Asides the mean, the range is also affected by outliers.

To explain the term outlier better, let us use numerical example.

**Example 1**: What is the outlier in the given text scores gotten by Oji in 7 of her courses at school? 50, 60, 1, 60, 68, 56, and 55.

Answer: Since 6 out of the 7 scores are within the range of 50-60, and only the one she got 1 is extremely different, then 1 is the outlier in the distribution.

Outliers may not necessary mean the least number or highest number, they are basically the values that are unusually large or small.

**SELF-ASSESSMENT EXERCISE**

Identify the outlier(s) in the given distribution.
120, 123, 1, 140, 650, 134, 122, 145 and 128.

## 4.0 CONCLUSION

The unit focuses on the measures of central tendency and dispersion. Seven of these measures were listed and discussed. From what was discussed, it was clear that the commonly used measures of central tendency are mean, median and mode with their primary objective being

to measure the central location of a distribution or data set. The measures of dispersion on the other hand measures the deviation of data set from the center with range, mean deviation, standard deviation and variance discussed in the unit. The weaknesses and strength of the measures were highlighted and the unit ended with the discussion of the outlier which often affects the mean and the range.

## 5.0 SUMMARY

In summary, the under listed were discussed in the unit.

1. The most common measures of central tendency are the mean, median and mode.
2. The range, variance, mean deviation and standard deviation are measures of dispersion and they measure the deviation of data set from the mean.
3. The calculation of these measures were explained with hypothetical examples.

## 6.0 TUTOR-MARKED ASSIGNMENT

1. Consider a sample with data values of 16, 20, 12, 17, 18 and 16. Compute the mean, mode and median.
2. Identify the outlier in this data set: 20. 120, 33, 45, 25, 46.
3. Find the variance and the standard deviation of the data set given below

| X | 10 | 11 | 13 | 15 | 17 | 14 | 12 |
|---|---|---|---|---|---|---|---|
| Frequency | 2 | 3 | 4 | 3 | 2 | 4 | 5 |

## 7.0 REFERENCES/FURTHER READING

Anderson, D. R., Sweeney, D. J. & Williams T. A. (2008). *Statistics for Business and Economics*, Tenth Edition. USA: Thomson Corporation.

Cramer D. & Howitt D. (2004). *The SAGE Dictionary of Statistics a Practical Resource for Students in the Social Sciences.* London*:* SAGE Publications.

Everitt, B. S. &Skrondal A. (2010). *The Cambridge Dictionary of Statistics*Fourth Edition. London: Cambridge University Press.

Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimmer, H., (2014). *Introduction to Statistics*. Rice University, Houston,

TX.Accessed:http://onlinestatbook.com/Online_Statistics_Education.pdf

Levine, D. M. & Stephan, D. F. (2005). *Even You Can Learn Statistics. A Guide for Everyone Who Has Ever Been Afraid of Statistics*. Pearson Education, Inc. Publishing as Pearson Prentice Hall Upper Saddle River, NJ. USA Accessed:file:///C:/Users/1/Downloads/[David_M._Levine,_David_F._Stephan]_Even_you_can_l(z-lib.org)%20(1).pdf.

National Bureau of Statistics (2018). *National Bureau of Statistics 2017 Demographic Statistics:* Bulletin. https://nigerianstat.gov.ng/download/775.

Obikeze D. S (1986). *Introductory Statistics for the Social Sciences.* Enugu: Fourth Dimension Publishing Co. Ltd.

Shafer, D. S. & Zhang, Z. (2012). *Introductory Statistics*. Flat World Knowledge. Washington, DC, USA.

Tokunaga H. T. (2016). *Fundamental Statistics for the Social and Behavioral Sciences*. California: SAGE Publications.

Watkins, J. C. (2016).*An Introduction to the Science of Statistics*: From Theory to Implementation (Preliminary Edition). Accessed: https://www.math.arizona.edu/~jwatkins/statbook.pdf.

.

# UNIT 2 DATA COLLECTION AND PRESENTATION

## CONTENTS



## 1.0 INTRODUCTION

Statistics as a course is all about data. This shows the importance of this unit to students. This unit explains all about data as they will ultimately be employed in statistical analysis designed to test the research hypothesis. This unit thus describes why researchers examine data and how data may be examined using tables, charts and graphs.

## 2.0 OBJECTIVES

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

- define data
- discuss types and classes of data
- explain the data collection methods
- explain how to present your data.

## 3.0 MAIN CONTENT

## 3.1 Data Collection and Presentation

As earlier mentioned, Statistics as a course is all about data. This shows how important this unit is to you. This unit explains all about data as they will ultimately be employed in statistical analysis designed to test the research hypothesis. So in this unit, you will be exposed to why researchers examine data and how data may be examined using tables, charts and graphs.

## 3.2 Data Collection

Data are statistical collection of quantities, characters, or symbols with related characteristics. For instance, the number of male students studying B.Sc Development Studies in NOUN in a given year- the number is called a data set. Data is a plural word and a single observation from the total data set is called a data point or datum.

### 3.2.1 Classification of Data

Data are classified into two major categories; Quantitative and Qualitative data.

**Quantitative data:** These are those enquiries whose values can be put into numerical figures. Thus, it deals with numbers and things you can measure objectively such as height, width, length, temperature and humidity, prices, area and volume etc. For example, the population of students in NOUN this year.

**Qualitative data:** These are those investigations whose values cannot be assigned numerical values. It deals with data that possess characteristics and descriptors that cannot be easily measured, but can be observed subjectively—such as smells, tastes, textures, attractiveness, and color.

**SELF-ASSESSMENT EXERCISE**

What do you understand by the term data?

### 3.2.2 Types of Data

There are two types of data; the primary and secondary data.
**Primary data:** Primary data are firsthand information collected by the researcher. The data so collected are pure and original and collected for a specific purpose. They have never undergone any statistical treatment

before. Primary data in most cases are customized to suit a particular analysis. The major weakness is that it is expensive. Primary data are sometimes referred to as raw information. An example of primary data is the census.

**Secondary data:** Secondary data are data that have been collected, compiled, published and made available in the form of record by some other organisation(s). In this case, the researcher does not need to go to the field to collect any data. Secondary data are impure in the sense that they have undergone been used by others. The major benefit of secondary that is that it is cheap to obtain. However, the chances of data errors are more likely than in primary data.

### 3.2.3 Methods of Primary Data Collection

1. *Personal investigation*: The person collects the data himself/herself by carrying out a personal study. The data so collected are reliable. This method is better suited for small projects.
2. *Observation*: This involves watching and counting events as they occur. This method is primarily used in traffic census, market survey etc. For instance, a researcher may wish to know the rate of usage of a public telephone line. He/she may decide to observe and record the number of persons using the line per day or hour as the case may be.
3. *Collection by researchers*: Here, trained researchers are employed to contact the respondents to collect data. This method is occasionally employed by some institutions such as the National Bureau of Statistics, the World Bank etc.
4. *Questionnaires*: Questionnaires are also used to ask specific questions that suit the study and get responses from the respondents. These questionnaires may be mailed as well.
5. *Telephone Interview*: The collection of data is done through asking questions over the telephone. This is normally less tedious and time saving.

### 3.2.4 Methods of Secondary Data Collection

1. Official publications such as the Ministry of Finance, Statistical Departments of the Government, Economic Boards, Government Agencies, etc. For instance, the Central Bank of Nigeria, National Bureau of Statistics, and the National Population Commission, do publish relevant statistical data.
2. Data published by Chambers of Commerce and trade associations and boards.
3. Articles in the newspaper and journals.

4. The internet

**SELF-ASSESSMENT EXERCISE**

Discuss the different kinds of data known to you.

## 3.3 Data Presentation

Data may be presented in form of tables, graphical presentations or charts.

### 3.3.1 Tables

A table is an orderly arrangement of data in rows and columns, or possibly in a more complex structure. Tables are very effective medium for the organisation and presentation of statistical information.

### 3.3.2 Uses of Tables

The following are the uses of tables:

1. To present data in a logically arranged and understandable form.
2. To identify peculiar features of the data as well as facilitate comparisons through row and column arrangements
3. To show the associations and patterns of relationships among variables
4. To facilitate speedy and easily decision taking since data are presented in understandable form

### 3.3.3 Characteristics of a Table

a) A table must be simple
b) A table must be easy to understand
c) A table must be numbered if they are more than one
d) A table must have a title
e) The sub headings for row and column must be stated
f) When a table contains secondary data, the source of the data must be shown
g) In a table, the unit of counting should be indicated
h) Items in a table must be arranged alphabetically, chronologically etc.

**Example of a Table**
**Records of students of NOUN secondary school**

| S/N | Name | Sex | Age | Height (m) | Weight (kg) |
|---|---|---|---|---|---|
| 1 | Peter Okafor | M | 14 | 1.56 | 46 |
| 2 | John Yinusa | M | 13 | 1.58 | 43 |
| 3 | Mary Ojo | F | 15 | 1.50 | 47 |
| 4 | Olufemi Adebayo | M | 14 | 1.61 | 50 |
| 5 | Vivian Peters | F | 16 | 1.54 | 50 |
| 6 | Jessica Akinfemi | F | 15 | 1.55 | 51 |
| 7 | Tunde Oyelami | M | 13 | 1.57 | 49 |
| 8 | AbubakarAudu | M | 14 | 1.60 | 52 |
| 9 | Hauwa Bello | F | 15 | 1.55 | 51 |
| 10 | Victor Eze | M | 16 | 1.58 | 53 |

**Source: NOUN Student Record, 2017**

**SELF-ASSESSMENT EXERCISE**

What is a table? Explain its uses in statistics.

## 3.3.4 Frequency Distribution

Data are only useful when they have been organised and ordered in a meaningful manner. Data can be presented in various forms depending on the type of data collected. A frequency distribution is a form of data presentation in which data are presented either in a graphical or tabular format, that displays the frequency of various outcomes in a sample within a given interval. It is thus a table showing how often each value (or set of values) of the variable in question occurs in a data set. A frequency table is used to summarise categorical or numerical data. Frequencies are also presented as relative frequencies, that is, the percentage of the total number in the sample.

## 3.3.5 Frequency Distribution for Ungrouped Data

Ungrouped data are data given as individual data points. E.g., 2. 3, 5, 6, 7.

**Example**
The test scores of 20 students are given as follows: 23, 26, 11, 18, 09, 21, 23, 30, 22, 11, 21, 20, 11, 13, 23, 11, 29, 25, 26, 26.
Note that the term frequency refers to the number of times an observation occurs or appears in a data set. Hence, in case of repetitions, the frequency increases. The table below will help you understand this better.

| Mark obtained in test | Tally | Frequency (No. of students) |
|---|---|---|
| 9 | / | 1 |
| 11 | //// | 4 |
| 13 | / | 1 |
| 18 | / | 1 |
| 20 | / | 1 |
| 21 | // | 2 |
| 22 | / | 1 |
| 23 | /// | 3 |
| 25 | / | 1 |
| 26 | /// | 3 |
| 29 | / | 1 |
| 30 | / | 1 |
| Total | | 20 |

## 3.3.6 Frequency Distribution for Grouped Data

In statistics, we often deal with large numbers in which case frequency distribution as presented above becomes hectic. To overcome this, numbers are grouped into categories called class intervals. Therefore, grouped data refers to data given in intervals. While lass interval refers the range of values incorporated within a given group.

**Properties of Class Interval**

a. Each interval has both upper and lower boundaries. The upper boundary is defined as the highest and the lower boundary is fixed by the lowest integer or whole number. E.g., the interval 7 to 18 has 18 as the upper boundaries and 7 as the lower boundaries
b. Each interval has an upper and lower limit
c. Each interval has a midpoint
d. Each interval has a size

**Example**

The exam scores of 37 students are presented below:
32, 52, 93, 50, 75, 79, 79, 74, 78, 61, 72, 72, 90, 54, 90, 84, 57, 58, 63, 91, 92, 67, 70, 68, 71, 72, 91, 73, 73, 91, 78, 77, 75, 80, 81, 83, 84.

The first step is to rearrange the list in an ascending order.
32, 50, 52, 54, 57, 58, 61, 63, 67, 68, 70, 71, 72, 72, 72, 73, 74, 75, 75, 77, 78, 79, 79, 80, 81, 83, 84, 84, 90, 90, 91, 91, 92, 93.

Next, we select appropriate class interval for grouping them. Thereafter, we tally them and the frequencies for each category will emerge. Here we have selected a group of ten. i.e. 0 -9, 10 – 19, 20 -29 etc.

The frequency distribution for the above data using a class interval of ten is illustrated below:

| Class Interval | Tally | Frequency |
|---|---|---|
| 30-39 | / | 1 |
| 40-49 | - | 0 |
| 50-59 | //// | 5 |
| 60-69 | //// | 4 |
| 70-79 | //// | 15 |
| 80-89 | //// //// //// | 5 |
| 90-99 | //// // | 7 |
| Total | | ∑f or N = 37 |

**SELF-ASSESSMENT EXERCISE**

i. What is a frequency distribution?
ii. Differentiate betweenfrequency distribution for grouped and ungrouped data.

## 3.4 Graphical Presentation

Frequency distributions are sometimes presented in graphical forms. There are several forms of graphical presentation. However, the common ones include histograms, polygons, line graphs, ogive etc.

### 3.4.1 Histogram

A histogram is a graphical display of statistical information that uses rectangular bars of different heights to show the frequency of data items in successive numerical intervals of equal size. In the most common form of histogram, taller bars show that more data fall in that range with the independent variable plotted along the horizontal axis and the dependent variable plotted along the vertical axis. The data appear as colored or shaded rectangles of variable area.

The illustration below, is a histogram showing the results of students TMA for a course.

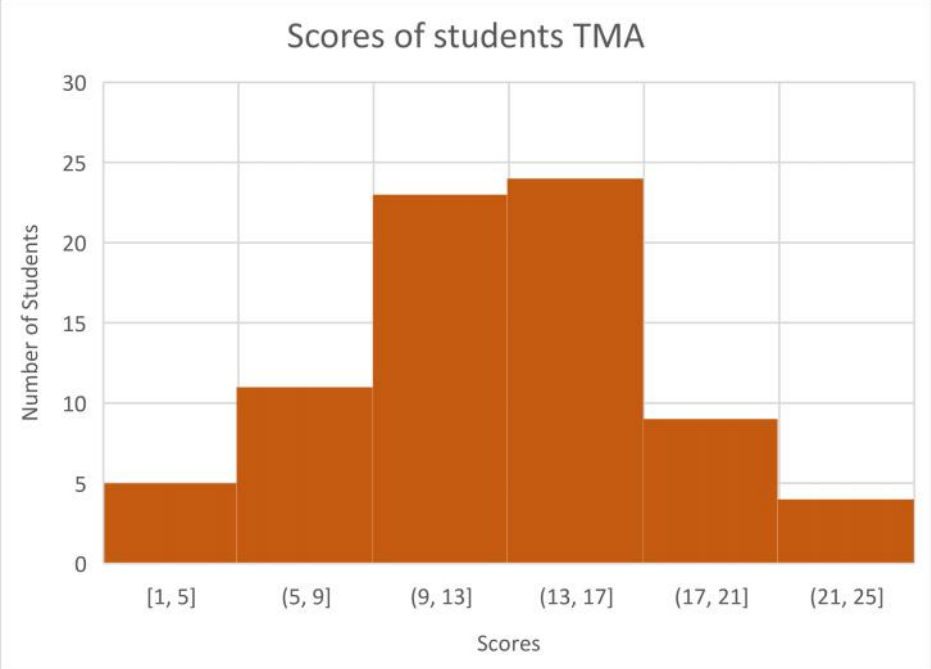

Scores of students TMA
30
25
20
15
10
5
0
Number of Students
[1, 5]
(5, 9]
(9, 13]
(13, 17]
(17, 21]
(21, 25]
Scores

***Fig. 2.1***

### 3.4.2 Polygon

The frequency polygon is a graphical representation of frequency distribution. It is similar to the histogram and is obtained by plotting class frequency as ordinates against the centre point (class mark) of class interval. It comprises of a series of line linking the midpoints of the tops of a rectangular bar.

Example of a polygon is shown below.

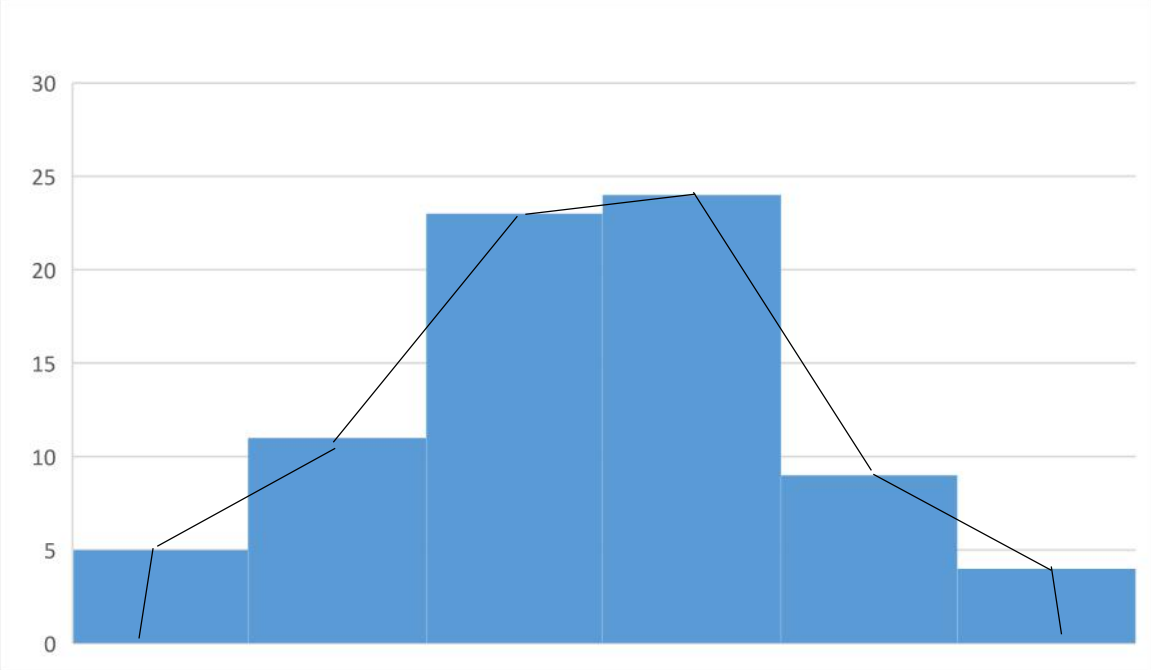



***Fig.2.2***

### 3.4.3 Statistical Bar Chart

A bar chart is a representation of figures or data using rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a line graph. Bar chart can be a simple chart or a component chart.

**Simple bar chart**: This is a bar chart that is made up of a unique components and a number of unconnected bars with the height representing the values of the respective categories.

**Example of a bar chart**

Average Rainfall for 3rd-7th May 1970

**Component bar chart:** This is a bar chart that is divided into a number of segments and partition with the height representing the values of a component of the category it represents. It is suitable for presenting a data where each category is made up of two or more components, parts or groups.

**Example of Component bar chart**

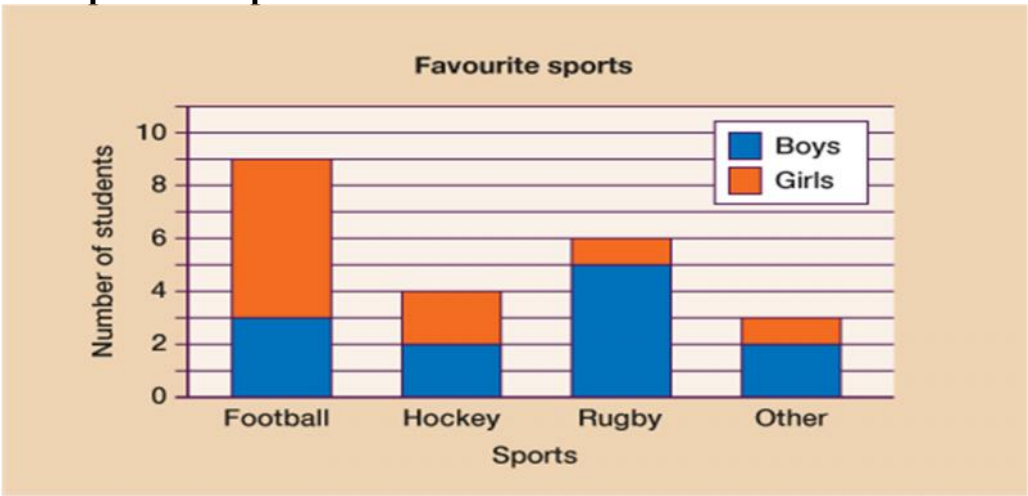



***Fig.2.3***

**Note that we have two groups or categories here; Boys and Girls**

## 3.4.4 Pie Chart

This consists of using a circle to represent the relative sizes of various categories of frequency values. In constructing a pie chart, the frequencies of all the categories are converted into degrees by relating them to 360 degree (360º).

Thus a 20% frequency becomes $\frac{20}{100} X \frac{360}{1} = 72^{o}$.

Using a protactorfrom a mathematical set, the degrees are marked out along the circumference of the circle. The pie chart is most suitable for representing categorical data.

**Example of a pie chart**

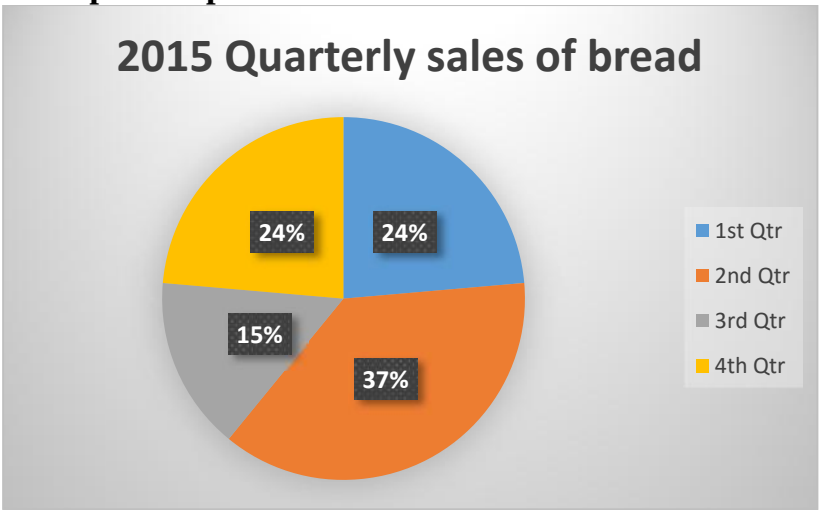



***Fig.2.4***

Note that an addition of the categories must give you 100% i.e. (24+24+37+15) = 100%.

**SELF-ASSESSMENT EXERCISE**

List the forms of graphical representation of frequency distributions you know.

## 4.0 CONCLUSION

This unit has been able to explain the meaning of data, types of data and method of data collection. It also explains the methods of data presentation. The unit thus succeeds in showing how students and researchers can present statistical data during research.

## 5.0 SUMMARY

At the end of this unit, students are expected to know the meaning of data, types of data and the procedures for generating the data they require for their research. Equally important is the fact that any enterprising student should be concerned with how to use them to achieve relevant results.

## 6.0 TUTOR MARKED ASSIGNMENT

1. What is data? Discuss the methods of data collection
2. The data obtained by measuring the age of 20 randomly selected students enrolled in freshman courses at a university are listed below. Present the data in a frequency table.
   18, 19, 18, 18,19, 24,19, 18,19, 20, 18, 18, 22, 21, 20, 20, 18, 17, 18 19

3. Discuss the following graphical presentations
   - ✓ Histogram
   - ✓ Polygon
   - ✓ Simple Bar Chart:
   - ✓ Component Bar Chart
   - ✓ Pie Chart

## 7.0 REFERENCES/FURTHER READING

Esan F.O. and Okafor, R.O. (2010): *Basis Statistical Methods (revised edition)* Lago: Toniichristo Concept

Gupta, C. B. (1983). *An Introduction to Statistical Methods.* New Delhi: Vikas Publishing House PVT Ltd

Obikeze D. S (1986). *Introductory Statistics for the Social Sciences*. Enugu: Fourth Dimension Publishing Co. Ltd.

# UNIT 3 PROBABILITY

## CONTENTS



## 1.0 INTRODUCTION

The aim of this unit is to introduce you to the concept of probability distribution, its discussion, calculation and interpretation of result.

## 2.0 OBJECTIVES

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

- explain probability theory
- differentiate between discrete and continuous probability distribution
- discuss how to use binomial distribution formula to solve probability problems
- discuss how to use distribution formula to solve probability problems
- Identify which of the two distribution functions is to be used to solve a given problem.

## 3.0 MAIN CONTENT

### 3.1 Probability Theory

The probability theory began in the seventeenth century in France when two great French mathematicians, Blaise Pascal and Pierre de Fermat, started a correspondence over the games of chance. Today, the

probability theory is a well-established and recognized branch of mathematics with applications in most areas of science and engineering

### 3.1.1 What is Probability?

The Probability of an event expresses the likelihood of the event occurring. Probability result ranges from 0 to 1. If it is 0, it means for certainty, the event cannot occur. On the other hand, if the answer is 1, it means it is certain the event will occur. The higher the probability of an event, the more likely it is that the event will occur.
If we denote probability with P, then,

$$\text{P (event)} = \frac{No\ of\ favourable\ outcomes}{No\ of\ all\ possible\ outcomes}$$

**SELF-ASSESSMENT EXERCISE**

What is probability?

## 3.2 Discrete Probability

In the toss of a fair coin, since the coin is fair, the two possible outcomes ("heads" and "tails") are both equally probable. The probability of "heads" equals the probability of "tails"; and since no other outcomes are possible, the probability of either "heads" or "tails" is 1/2 (which could also be written as 0.5 or 50%). In rolling of a die once, the possible outcomes is 6, that is a 1,2,3,4,5, or 6 may come out, if the question says ‘what is the probability of having a 2? Note that the number of favourable outcome is 2 while the number of possible outcomes is 6, therefore the answer is 1/6.

**Example 1:** If 3 coins are flipped at the same time, find the probability of

(i) obtaining 3 heads
(ii) 2 tails.

**Solution**
**Discrete Probability Distribution (Possible outcomes)**

| | |
|---|---|
| HHH | 3 heads |
| HHT | 2 heads 1 tail |
| HTH | 1 head 1 tail and one head |
| HTT | 1 head 2 tails |
| THH | 1 tail 2 heads |
| THT | 1 tail 1 head 1tail |
| TTH | 2 tail 1 head |
| TTT | 3 tails |

There are 8 possible outcomes, the probability of obtaining 3 heads is equal to 1/8.
The probability of obtaining 2 tail is 3/8.

**Example 2**: Find the probability of rolling doubles on two six-sided dice numbered from 1 to 6.
The first step is to find all the total possible outcomes or the sample space.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1,1 | 1,2 | 1,3 | 1,4 | 1,5 | 1,6 |
| 2 | 2,1 | 2,2 | 2,3 | 2,4 | 2,5 | 2,6 |
| 3 | 3,1 | 3,2 | 3,3 | 3,4 | 3,5 | 3,6 |
| 4 | 4,1 | 4,2 | 4,3 | 4,4 | 4,5 | 4,6 |
| 5 | 5,1 | 5,2 | 5,3 | 5,4 | 5,5 | 5,6 |
| 6 | 6,1 | 6,2 | 6,3 | 6,4 | 6,5 | 6,6 |

From the table above, the number of possible outcomes is 36 while the number of favourable outcomes (that is rolling doubles on two six-sided dice) is 6, that is (1,1; 2,2; 3,3; 4,4; 5,5; 6,6) the answer will be equal to $\frac{6}{36} = \frac{1}{6}$.
We can also find the probability of picking a number of an item/object from a group of items/object. For instance, one can find the probability of picking a red ball out of a number of red balls and blue balls.

**Example 3:**
A box contains 3 red balls and 4 blue balls. What is the probability of picking 2 red balls?

Note that each of the red balls has equal chance of being selected. Since there are 3 red balls, the number of favourable outcomes is 3 while the number of possible outcomes is 7 (that is 3 red balls plus 4 blue balls). Therefore, the answer is 3/7 or approximately 0.4.

An item can be chosen with or without a replacement, in this case the method of selection will be different. If items/objects are chosen with replacement, the total number of possible outcomes remains the same, however, if they are chosen without replacement the number of possible outcomes reduces. For example: A basket contains 4 green balls, 3 black balls and 2 white balls, if a ball is selected with replacement, what is the probability of selecting (i) a green ball? (ii) a black ball (iii) a green ball and a white ball without replacement.

**Solution:**
Note that there are 9 balls altogether

i. Since there are 4 green balls, then 4/9 balls will be selected with replacement,
ii. Since there are 3 black balls, then 3/9 or 1/3 balls will be selected with replacement,
iii. A green ball and a white ball without replacement:
Let's assume at first a green ball is selected,

$= \frac{4}{9} \times \frac{2}{8} = \frac{4}{9} \times \frac{1}{4} = \frac{1}{9}$

And then a white ball is selected,

$= \frac{2}{9} \times \frac{4}{8} = \frac{2}{9} \times \frac{1}{2} = \frac{1}{9}$

$= \frac{1}{9} + \frac{1}{9} = \frac{2}{9} =$ **0.2222 = <u>22.22%</u>**

OR

$$\frac{4}{9} \times \frac{2}{8} + \frac{2}{9} \times \frac{4}{8} = 0.2222$$

= 22.22%

## 3.2.1 Types of Discrete Probability Distribution

There are different classifications of probability distributions. They include the binomial distribution and Poisson distribution. The different probability distributions serve different purposes and represent different data generation processes.

**1. Binomial probability distribution**

The binomial distribution is a probability distribution that summarises the likelihood that a value will take one of two independent values under a given set of parameters or assumptions. The underlying assumptions of the binomial distribution are that there is only one outcome for each trial, that each trial has the same probability of success, and that each trial is mutually exclusive, or independent of each other.

The binomial distribution is a common discrete distribution used in statistics, as opposed to a continuous distribution, such as the normal distribution. This is because the binomial distribution only counts two states, typically represented as 1 (for a success) or 0 (for a failure) given a number of trials in the data. The binomial distribution, therefore, represents the probability for x successes in n trials, given a success probability p for each trial. It is important to note that the Binomial distribution is a discrete probability distribution. It is often used in social science statistics as a building block for models for dichotomous outcome variables, like whether APC or PDP win an upcoming election. Other examples include whether an individual will die within a specified period of time; or whether, in an examination, a student will choose a correct answer at random.

**Characteristics of Binomial Distribution**

i. There is a fixed number of n trials
ii. The trials are mutually independent
iii. There is a constant probability of success at each trial
iv. The variable is the total number of successes in n trials.

The formula for binomial distribution is given as:

$P(x) = (NC_K) = P^k q^{n-k}$

Where:

$k = 0, 1, 2, ..., n,$

$p = Probability\ of\ success\ or\ probability\ of\ the\ event\ occuring$

$q = 1 - p\ (Probability\ of\ failure)$

$nc_K = n\ combination\ k$ given as: $\frac{n!}{k!(n-k)!}$

x = number of value required

P = Probabilistic value

**Example**

A multiple choice question contains 20 questions with answer choice A, B, C and D. Only one answer choice to each question represents a correct answer. Find the probability that a student will answer exactly six questions correctly if he makes random guesses on all 20 questions.

**Solution:**

Using Binomial probability distribution formula:

$P(X) = (nc_K)\ P^k q^{n-k}$

$n = 20$

$k = 6$

$P = 0.25$

$q = 0.75$

$nc_K = \frac{n!}{k!(n-k!} = \frac{20!}{6!(20-6!)}$

$= \frac{20!}{6!(14!)} = \frac{20\times19\times18\times17\times16\times15\times(14!)}{6\times5\times4\times3\times2\times1\times(14!)}$

$\frac{27907200}{720} = 38{,}760$

$P(6) = 38{,}760\ (0.25^6 \times 0.75^{20-6})$

$= 0.1686$ or 16.86%

**2. Poisson Distribution**

In statistics, a Poisson distribution is a statistical distribution named after French mathematician Siméon Denis Poisson. It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time. It shows how many times an event is likely to occur within a specified period of time. It is used for independent events which occur at a constant rate within a given interval of time. It is a tool that helps to predict the probability of certain events

from happening when you know how often the event has occurred. The distribution gives the probability of a given number of events happening in a fixed interval of time. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

Suppose we are counting the number of occurrences of an even in a given unit of time, distance, area or volume, then, the Poisson distribution may be useful to model events such as:

- The numbers of times vehicular accidents occur in each day.
- The number of patients arriving in an emergency room between 10 and 11 pm
- The number of photons hitting a detector in a particular time interval

It is based on the assumption that:

✓ Events are occurring independently
✓ The probability that an even occurs in a given length of time does not change through time. In other words, the theoretical rate at which the events are occurring does not change through time.
✓ The occurrences must be uniformly distributed over the intervals being used.

More loosely, we may say that the events are occurring randomly and independently. Then X, the number of events in a fixed unit of time has a Poisson distribution.

The formula for the Poisson distribution is given as:

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

Where:

$\lambda$ = average number of events per interval

$e$ = is the number 2.71828 (Euler's number) that is the base of the natural logarithms.

$x$ = k takes values 0, 1, 2,....

$x! = k \times (k - 1) \times (k - 2) \times ... \times 2 \times 1$ *is the factorial of k*

The Poisson distribution is different from the binomial distribution in these two major ways:

➢ The binomial distribution is affected by sample size *n* and the probability $p$ whereas the Poisson distribution is affected only by the mean ($\mu$).

- In a binomial distribution, the possible values of the random variable $x$ are 0,1,2,3,….$n$ but a Poisson distribution has possible $x$ values of 0,1,2,3…., with no upper limit.

**Example**
If 530 droughts have occurred in the last 500 years, assuming this event is suitable in Poisson distribution, find the probability of 2 hurricanes occurring in a randomly selected year.

**Solution**
Find $\lambda$, the mean number of hurricane, therefore:

$$\lambda = \frac{number\ of\ hurricane}{100} = \frac{530}{100} = 5.3$$

$$P(2) = \frac{\lambda^2 e^{-\lambda}}{2!}$$

$$P(2) = \frac{5.3^2(2.71828)^{-5.3}}{2!}$$

$= 0.0701$

**Example 2**
A small business receives an average of 12 customers per day. What is the probability that the business will receive exactly 8 customers in one day?

$$\lambda = \mu = 12$$
$$e = 2.71828$$
$$P(8) = P(2) = \frac{12^8(2.71828)^{-12}}{8!}$$
$$= 0.0655$$
$$= 6.55\%$$

**SELF-ASSESSMENT EXERCISE**

i. Differentiate between binomial and Poisson probability distributions.

## 3.3 Continuous Probability

Continuous probability distribution is a type of distribution that deals with continuous types of data or random variables. The continuous random variables deal with different kinds of distributions. A typical example is the normal distribution. This will be discussed very shortly.

### 3.3.1 Normal Distribution

Normal distribution is one of the most useful continuous probability distributions in statistics. Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. It is a symmetric distribution where most of the observations cluster around

the central peak and the probabilities for values further away from the mean tapers towards both directions.

A normal distribution has a bell-shaped density curve described by its mean and standard deviation. The density curve is symmetrical, centered about its mean, with its spread determined by its standard deviation. The normal distribution is a bell-shaped frequency distribution, symmetric, unimodal, and asymptotic. The mean, median, and mode are all equal.
The probability that a normally distributed variable X with known μ and $\sigma$ is in a particular set, can be calculated by using the fact that the fraction Z = $\frac{x-\mu}{\sigma}$ has a standard normal distribution.

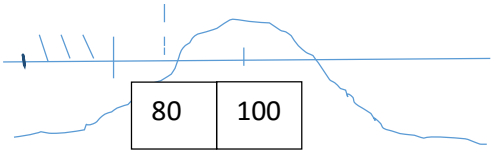
**Characteristics of a Normal Distribution**

(i) The graph is symmetrical about the y-axis
(ii) The range of x is from $-\infty$ $to$ $+\infty$
(iii) The x-axis is asymptotic to the curve at both extremes.
(iv) The curve has a maximum at origin.

**Example**
Normally distributed IQ score have a mean of 100 and standard deviation of 15. Use the standard z-table to answer the following questions. What is the probability of randomly selecting someone with an IQ score that is (a) less than 80, (b) greater than 136.

(a)

$Z = \frac{x-\mu}{\sigma}$
$\mu = 100$
$\sigma = 15$
$x = 80$
$Z = \frac{80-100}{15} = -1.33$




P (x < 80)
Check the negative Z table, find the value of area that correspond to the z value.
= 0.09176 or 0.09176 x 100
= 9.176%
(2)
$\mu = 100$
$\sigma = 15$
$x = 130$
P (x > 130)
$Z = \frac{136-100}{15} = 2.4$
Using the positive Z table, check the value of 2.4

= 0.9918
P(x > 136) =1- p(x < 136)
= 1- 0.9918
**= 0.0082 or 0.82%**
Therefore, the probability of randomly selecting someone with an IQ score that is
greater than 136 is 0.0082 0r 0.82%.

**SELF-ASSESSMENT EXERCISE**

i. What is a continuous probability distribution?
ii. What are the characteristics of a normal distribution?

## 4.0 CONCLUSION

From our discussions, we have learnt and become familiar with the concept of probability and its distributions such as binomial, poisson and normal distribution.

## 5.0 SUMMARY

In summary, it is expected that having finished this unit, you should have become familiar with the probability theory. You are also expected to have understood when/where to apply discrete and continuous probability distributions given any question that involves knowledge of probability.

## 6.0 TUTOR-MARKED ASSIGNMENT

A study shows that 35% of the people entering a store make a purchase. Using a binomial distribution and Poisson distribution find the probability that out of 25 people entering the store 8 or more will make a purchase

## 7.0 REFERENCES/ FURTHER READING

Adedayo, A.O. (2006). *Understanding Statistics.* Lagos: JAS Publishers.

Dominick, S. & Derrick, R. (2011). *Statistics and Econometrics.* New York: McGraw Hill.

Grinstead& Snell's (2006). *Introduction to Probability*. The CHANCE Project1, Version dated 4 July 2006.

Loto, M.A., Ademola, A.A., & Toluwase, J.A. (2008). Statistics Made Easy. Pub., Concept Publications LTD. Lagos.

Webster's Revised Unabridged Dictionary. G & C Merriam, 1913, Probability

# MODULE 3 INFERENTIAL STATISTICS



## UNIT 1 SAMPLING

**CONTENTS**



## 1.0 INTRODUCTION

In order to make statistical work meaningful, statisticians generalise from what they find in the figure at hand to the wider phenomenon which they represent through data. These set of data are regarded as a sample drawn from a larger “universe”. We analyse the data of the sample in order to draw conclusion about the corresponding universe or population. This unit sets out to discuss and understand the topic “sampling”.

## 2.0 OBJECTIVES

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

- discuss Sampling
- discuss Sample size
- explain Population
- describe types and method of sampling
- discuss Sampling with and without replacement etc.

## 3.0 MAIN CONTENT

## 3.1 Basic Concepts

**POPULATION**
Population refers to the total units of individuals whose characteristics are under study. It deals with the aggregation of all cases such as - persons, objects, etc. Thus in a study of students, the population may be defined as

- Persons registered for B.Sc. Development Studies in NOUN
- Persons enrolled into NOUN in Nigeria

The first definition deals with students who are registered for B.Sc. Development Studies and doesn't include other programmes while the second definition is broader and captures all students in NOUN irrespective of the programme.

Population can be finite or infinite. A finite population is a population that contains finite number of units while an infinite population is composed of an infinitely large number of elements. An infinite population can at best be imagined

## 3.2 Sampling

Sampling in statistics is a systematic way of selecting a part or subset from the total population so as to make inferences from them. It therefore involves the process and technique employed in selecting a representative part of a population for the purpose of determining parameters or characteristics of the whole population.

**SELF-ASSESSMENT EXERCISE**

Define the term, sampling.

## Sample

A sample is subset of the population. It is a smaller, controllable form of a large population and contains all characteristics of the population. It is used in statistical testing when population sizes are too large for the test to include all possible members or observations and as such, represent the population as a whole and not reflect bias toward a specific attribute. Samples are used in a variety of settings where research is conducted. Scientists, marketers, government agencies, economists, and research groups are among those who use samples for their studies and measurements. For instance, a researcher may wish to know how many students studied for less than 30 hours for TMA test in NOUN.

Since the population of NOUN students is over 400,000 with all always taking part in the TMA test, reaching out to each and every student may be extremely difficult and time consuming. In fact, by the time the data from the population have been collected and analyzed, a couple of years would have passed, making the analysis worthless since a new population would have emerged. What should be done is to take a sample of the population (say 200) and get data from this sample. This selection is done in a random manner to give everyone from the population an equal and likely chance of being added to the sampled group.

### Sample Size

A sample size is a part of the population chosen for a survey or experiment. For example, you might take a survey of car owners' brand preferences. You may not wish to survey *all* the millions of car owners in the country due to the bulky and unending nature of embarking on such task, so you take a sample size. That may be several thousand owners. The sample size is *a* representation of all car owners' brand preferences. If you choose your sample wisely, it will be a good representation. The size of a study sample has important economic and statistical implication. Some statistical methods require large sample for proper application while others are suited for small samples. What constitute a large or small sample depends on the size of population, nature of study and orientations of the user. Conventionally, a sample size is regarded as small when it is less than 30 and large when it is up to 50 or more.

### SELF-ASSESSMENT EXERCISE

What do you understand by the terms sample and sample size?

## 3.3 Types and Method of Sampling

Generally, sampling method is divided into two broad types; Non-Probability and Probability Sampling also known as biased and random sampling.

### 3.3.1 Non-Probability Sampling

This is a sampling technique in which the researcher selects samples based on the subjective judgment rather than random selection. It is a method in which it is not possible to determine the probability of each element being included in the sample. In non-probability sampling, not all members of the population have a chance of being selected in the study. This means that the likelihood of being included in the sample differs from one population element to another; certain elements or groups have advantage over others. It is used in studies where it is not possible to draw random probability sampling due to time or cost considerations or for other reasons. It is based on other methods of observation and it is widely used in qualitative research.

### 3.3.2 Types of Non-Probability Sampling

1 **Haphazard or Accidental Sampling**
Also known as Convenience Sampling, it is a non-probability sampling technique where samples are selected from the population only because they are conveniently available to the researcher. These samples are selected only because they are easy to recruit and the researcher does not consider selecting the samples that represent the entire population. Such samples are biased because the researcher may deliberately approach some kinds of respondents and avoid others.

2. **Consecutive Sampling**
This non-probability sampling technique is very similar to convenience sampling, with a slight variation. Here, the researcher picks a single person or a group of samples, conducts research over a period of time, analyzes the results and then moves on to another subject or group of subject if needed.

**3. Purposive Sampling**
This is the sampling method whereby the investigator merely handpicks those cases considered to be typical or which are likely to possess the desired set of information or characteristics for inclusion in the sample. This is used primarily when there is a limited number of people that have expertise in the area being

researched, or when the interest of the research is on a specific field or a small group. A typical example of purposive sampling is the so-called quota sampling.

**Quota sampling:** Quota sampling involves selecting a sample to ensure that various subgroups in the population are represented in the sample in the same proportion.

**SELF-ASSESSMENT EXERCISE**

What are the types of non-probability sampling known to you?

### 3.3.3 Probability Sampling

A probability sampling method is any method of sampling that utilizes some form of random selection. In order to have a random selection method, you must set up some process or procedure to ensure that every element of the population gets an equal chance to be part of the selected sample. Probability sampling uses randomisation technique and is alternatively known as random sampling.

The various types of probability sampling include:

1. **Simple Random Sampling**
   A simple random sampling is a method of selecting samples where every unit in the population has a known probability of being selected. The probabilities of selection are equal and sample is obtained by making a list of the population units, assigning numbers to the individuals (sample) and then randomly choosing from those numbers through an automated process either by employing a computer or balloting. Finally, the numbers that are chosen are the members that are included in the sample. Simple random sampling is simple to accomplish and is easy to explain to others. Because simple random sampling is a fair way to select a sample, it is reasonable to generalise the results from the sample back to the population. Simple random sampling is not the most statistically efficient method of sampling and you may, just because of the luck of the draw, not get good representation of subgroups in a population. To deal with these issues, we have to turn to other sampling methods.

2. **Stratified Random Sampling**
   Stratified Random Sampling is a method of sampling that involves the division of a population into smaller homogeneous subgroups known as strata. In stratified random sampling or stratification, the strata are formed based on members' shared

attributes or characteristics. For instance, if the population is given as N, the population is divided into strata using features such as age, sex, size, geographical location N1, N2, N3, ... Ni, such that the combination of the strata becomes N1 + N2 + N3 + ... + Ni = N. Then a simple random sampling is done on each stratum. Stratified sampling is preferred over simple random sampling for some reasons which include:

Firstly, it assures that you will be able to represent not only the overall population, but also key subgroups of the population, Secondly, stratified random sampling will generally have more statistical precision than simple random sampling. This is so because due to the homogeneity of the strata, the variability within groups is lower than the variability for the population as a whole.

**3. Systematic Random Sampling**

Systematic sampling is a probability sampling method where the elements are chosen from a target population by selecting a random starting point and selecting other members after a fixed 'sampling interval'. For example, we may decide to sample every 20th or 65th member of the population. Systematic random sampling is a probability sampling only if the population elements are randomly distributed and the first case to be sampled is chosen by simple random sampling process. In systematic sampling, sampling interval is calculated by dividing the entire population size by the desired sample size.

The systematic sample is a variant of the simple random sample. Rather than referring to random number tables to select the cases that will be included in your sample, you select units directly from the sample frame. The procedure involved in systematic random sampling is very easy and can be done manually. The results are representative of the population unless certain characteristics of the population are repeated for every $n^{th}$ individual, which is highly unlikely. The process of obtaining the systematic sample is much like an arithmetic progression.

**4. Cluster Sampling**

A cluster is a group of units crowding together in the sample area or neighborhood. For example, a school is a cluster of students. Cluster sampling methods are modifications of stratified random sampling involving series of sampling processes at various levels and subgroups of the population. A cluster sampling is therefore a sampling method in which multiple clusters of people are created from a population where they are indicative of homogeneous

characteristics and have an equal chance of being a part of the sample. With cluster sampling, the researcher divides the population into separate groups, called clusters. Then, a simple random sample of clusters is selected from the population. The elements in each cluster are then sampled. If all elements in each sampled cluster are sampled, then it is referred to as a "one-stage" cluster sampling plan. If a simple random subsample of elements is selected within each of these groups, this is referred to as a "two-stage" cluster sampling plan. The researcher then conducts the analysis on data from the sampled clusters.

**SELF-ASSESSMENT EXERCISE**

Discuss any two kinds of probability sampling.

### 3.3.4 Advantages of Probability Sampling

The following are the advantages of probability sampling:

1. Probability sampling does not depend upon the existence of detailed information about the universe for its effectiveness.
2. Probability sampling provides estimates which have measurable precision.
3. It is possible to evaluate the relative efficiency of various sample designs only when probability sampling is used.

### 3.3.5 Limitations of Probability Sampling

Probability Sampling has some limitations. These limitations include:

1. Probability sampling requires a very high level of skill and experience for its use.
2. It requires a lot of time to plan and execute a probability sample.
3. The costs involved in probability sampling are generally large.

**SELF-ASSESSMENT EXERCISE**

What are the advantages and limitations of probability sampling?

## 3.6 Sampling with and without Replacement

When a sample is drawn from a population, be it a probability or non-probability sample, one may put back a selected case into the fold before drawing the next sample. One may on the other hand take out a chosen case from the population before making the next draw. In the first instance, the total population (N) remains the same at each draw while in

the second, the population decreases progressively by one at each subsequent drawing. The first method is referred to as sampling with replacement while the second is sampling without replacement.

Sampling is thus called with replacement when a unit selected at random from the population is returned to the population and then a second element is selected at random. Whenever a unit is selected, the population contains all the same units, so a unit may be selected more than once and can be chosen again and again. There is no change at all in the size of the population at any stage.

Sampling without replacement from a finite population means that the population will get exhausted after several samplings while, sampling with replacement from the same population means that no matter the number of samples taken, the population remains the same. Therefore, sampling with replacement turns a finite population into an infinite one. In a case of a large population, sampling with and without replacements produces similar results, but the case is different when the population is small. In the case of sampling with replacement, the possible number of samples of size "***r***"that can be drawn from population "**N**"is given as $N^r$, while in the case of sampling without replacement, it is given as $\frac{N!}{r!(N-r)!}$

**SELF-ASSESSMENT EXERCISE**

What do you understand by sampling with and without replacement?

## 3.6.1 Sampling Error

A sampling error is a statistical error that occurs when the subset of the population (sample) deviates from the true characteristics, attributes and behavior of the total population. This happens when a researcher fails to select a sample that represents the entire population such that the result obtained does not reflect and represent the results that would be obtained from the entire population.

Basically, there are two types of Sampling Errors:

*Biased errors***:** These errors arise from any bias in selection, estimation, etc. it is based on the personal prejudice or bias of the researcher. For instance, a researcher is required to collect a sample for a study using the simple random sampling but instead chooses other sampling method due to his/her bias. Bias may arise due to:

(i) in appropriate process of selection
(ii) in appropriate work during the collection; and
(iii) in appropriate methods of analysis

*Unbiased Errors***:** The Unbiased errors arise due to chance. The researcher may have not intentionally tampered with the sample and that the difference between the population and sample has occurred by chance. Despite carefully selecting a sample, the sampling error may occur because the subjects drawn from the population have individual differences. And therefore, the researcher must keep in mind that only the subset of the population is selected, and hence there will be a difference between the population and a sample.

**SELF-ASSESSMENT EXERCISE**

What are the types of sampling error in statistics?

## 3.6.2 Merits of Sampling

1. Less time consuming. Since the sample is a study of a part of the population, considerable time and labour are saved when a sample survey is carried out.
2. Less cost. Although the amount of effort and expense involved in collecting information is always greater per unit of the sample than a complete census, the total financial burden of a sample survey is generally less than that of a complete census. This is because of the fact that in sampling, we study only a part of population
3. More reliable results. Although sampling involves certain inaccuracies owing to sampling errors, the result obtained is generally more reliable than obtained from a complete count. There are several reasons for this. First, it is always possible to determine the extent of sampling errors. Secondly, other types of errors to which a survey is subject, such as inaccuracy of information, incompleteness of returns, etc., are likely to be more serious in a complete census than in sample survey
4 More detailed information is gotten. Since sampling saves time and money, it is possible to collect more detailed information in a sample survey.
5. Sampling method is the only method that can, be used in certain cases. There are some cases in which the census method is inapplicable and the only practicable means is provided by the sample method.

### 3.6.3 Limitations of Sampling

Some of the difficulties involved in sampling include:

1. In some cases, results obtained may be inaccurate and misleading leading to sampling errors. This is so if sampling is not carefully planned and executed as sampling procedure is not perfect.
2. It requires the service of experts, such that in the absence of qualified and experienced persons, the information obtained from sample surveys cannot be relied upon.
3. If the information is required for each and every unit in the domain of study, a complete enumeration survey is necessary.
4. There exist chances of biasness since the choice of sampling method is a judgmental task.
5. Improper selection of sampling techniques may cause the whole process to be wrong.
6 Selection of proper sample size is difficult.
7. Sampling may exclude some data that may not be homogenous to the data that are taken. This affects the level of accuracy in the results.

## 3.7 Sampling Distribution

Sampling distribution is the distribution of a given statistic among all possible samples of a given size that can be drawn randomly from the population. It is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population. Observe that in this case, we are concerned with the frequency of the distribution, not of raw scores or variable attributes but of statistics such as mode, mean, median, variance, and standard deviation. Sampling distributions therefore is important because they provide a major simplification of statistical inferences.

**SELF-ASSESSMENT EXERCISE**

What is sampling distribution?

## 3.8 Statistical Estimation

Statistical estimation refers to the process of generalising or making inference from the observed sample values to population values which are unknown. It is the process of inferring the population parameter from a sample statistic.

### 3.8.1 Choosing the Best Estimator

An estimate refers to the value of the sample statistic which is taken as an approximation of the parameter value. It is the numeric value of the estimator of a given sample. An estimator on the other hand refers to the statistic which has been chosen to provide an estimate of the population value. It is the function of sample observations whose value at a given realisation of the observation gives the estimate of the population parameter. Thus, an estimator is a random variable calculated from the sample data that supplies either interval estimates or point estimates for population parameters.

The process of choosing a best estimator include the following:

**Bias**
An estimator is said to be biased, if the mean of its sampling distribution differs from the population mean. An estimator is said to be unbiased if the expected value of the estimator is equal to the population parameter being tested. A biased estimator produces a biased estimate and an unbiased estimator produces an unbiased estimate.

The sample mean $\bar{x}$ is an unbiased estimator of population mean. On the other hand, the sample variance $S^2$ is a biased estimator of the population variance $\sigma^2$. This is so since the mean of the sampling distribution of sample variance is not equal to population variance. Symbolically,

$$\frac{\sum(X - \bar{X})^2}{n} \neq \sigma^2$$

but

$$\frac{\sum(X - \bar{X})^2}{n} = \frac{N\sigma^2}{n-1} = \hat{\sigma}^2$$

Where $\hat{\sigma}^2$ is an unbiased estimate of population variance. By the same reasoning, sample standard deviation, S, is a biased estimator of population standard deviation.

### 3.8.2 Steps to Check Whether an Estimator is Unbiased or Not

a. Draw all possible samples of a given size from the population
b. Calculate the value of given estimator for all these samples separately
c. Take the average of all these values obtained in step b above.

If the average is equal to the population parameter, then the estimator is unbiased and if this average is more than the population parameter, then the estimator is said to be positively biased and when this average is less than the population parameter, it is said to be negatively biased.

**Efficiency**
An estimator is said to be the most efficient if with all possible alternatives, produces the least standard error. If two estimators have the same efficiency, the less biased is considered the better estimator.

**Consistency**
An estimator is said to be consistent if as the sample size increases, the estimates approaches the parameter, and the standard error correspondingly decreases. The sample mean thus is a consistent estimator of the population mean.

**SELF-ASSESSMENT EXERCISE**

Explain the steps involved in choosing a best estimator

## 4.0 CONCLUSION

In this unit, we have explained the meaning of sampling in statistics. We also explained its features, ways and the possibility of selecting a part or subset from the total population so as to make inferences with regards to the total population.

## 5.0 SUMMARY

This unit discusses sampling in some detail. The types and methods of sampling were also explained alongside the purpose of sampling and the relevance of sampling to statistics and research.

## 6.0 TUTOR-MARKED ASSIGNMENT

1. Differentiate between probability and non-probability sampling.
2. What is a statistical estimator?
3. Discuss the steps to check whether a given estimator is unbiased.

## 7.0 REFERENCES/FURTHER READING

Carlson, R. (2006). *A Concrete Introduction to Real Analysis*. CRC Press.

Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms*. OUP.

Grewal, P. S. (1990). *Methods of Statistical Analysis.* (2nd ed.). New Delhi: Sterling Publishers pvt. Ltd.,

Obikeze D. S (1986). *Introductory Statistics for the Social Sciences.* Enugu: Fourth Dimension Publishing Co. Ltd,

# UNIT 2 HYPOTHESIS AND SIGNIFICANCE TESTING

## CONTENTS



## 1.0 INTRODUCTION

In this unit, statistical hypothesis, its types and how to test them so as to determine the significance level will be discussed.

## 2.0 OBJECTIVES

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

- discuss hypothesis
- explain Type I and Type II errors
- explain Level of Significance
- describe One-tailed and two-tailed tests and their comparison
- describe hypothesis testing using Chi Square and t –test.

# 3.0 MAIN CONTENT

## 3.1 Hypothesis and Significance Testing

In statistical estimations, one may out of an idea, belief or previous knowledge make a conjectural statement, theoretical proposition or an assertion and may go on to draw a sample in order to test whether the proposition would be confirmed or disproved by the observed data. This form of inference is referred to as hypothesis testing. It is so because the propositions to be tested are stated in form of hypothesis. A hypothesis is therefore defined as a tentative proposition specifying some form of relationship between two or more variables made on the basis of limited evidence as a starting point for further investigation. The purpose of a hypothesis is to find the answer to a question. A formalised hypothesis will force us to think about what results we should look for in an experiment, and is usually made up of a dependent variable and an (or some) independent variable(s); the independent variables- the part of the experiment that can be changed and tested. The independent variable happens first and can be considered the cause of any changes in the outcome. The outcome is called the dependent variable.

**SELF-ASSESSMENT EXERCISE**

What is hypothesis testing?

## 3.2 Basic Concept

Basically, there are two types of hypothesis- substantive and null hypothesis

### 3.2.1 Substantive Hypothesis

The substantive hypothesis is also referred to as alternate hypothesis and represented by the symbol ($H_1$). It is a hypothesis on which a study is based. It makes a statement that suggests or advises that something is happening; a new theory is true instead of an old one and as such is the proposition which the researcher wishes to confirm using a data set. It is also known as the research hypothesis. It is always expressed in positive terms. Alternate hypotheses can be directional or non-directional. The directional hypothesis is a kind that explains the direction of the expected findings. This hypothesis in most cases is developed to examine the relationship among the variables rather than a comparison between the groups, and is tested with a two-tailed statistical test. The non-directional hypothesis is a kind that has no definite direction of the expected findings being specified. Examples include

- Research and advancement have impacted positively on economic growth in Nigeria
- There is a positive relationship between teenage pregnancy and the level of female drop out in secondary schools etc.

### 3.2.2 Null Hypothesis

The null hypothesis is a typical theory used in statistics which proposes that no statistical relationship and significance exists in a set of given variables, between two sets of observed data and measured phenomena. It is a proposition that undergoes verification to determine if it should be accepted or rejected in favor of an alternative proposition. Denoted by $H_0$, it is a logical converse of the alternative hypothesis and more or less a negation of the alternative hypothesis. It thus states the exact opposite of what a researcher predicts or expects by basically stating that there is no exact or actual relationship between the variables. Examples of null hypothesis include

- Violent video games have no impact on future acts of violence
- There is no significant relationship between foreign direct investment and economic growth in Nigeria
- Employees' job satisfaction is not positively related to their commitment to the organization

**SELF-ASSESSMENT EXERCISE**

What is the difference between null and alternative hypothesis?

### 3.2.3 Importance of a Hypothesis

i. It aims to encourage critical approach.
ii. It enables the researcher to develop a specific direction as well as better understanding about the subject matter of the study.
iii. It further assists in the careful and focused analysis of data collected.

### 3.2.4 Characteristics of a Good Hypothesis

i. A good hypothesis is always logical and affirmative. It is based on proper verification with clear and precise statement
ii. It should be understandable
iii. It should be testable and measurable
iv. It should contain a dependent and independent variable(s) and offer a balanced relationship between these variables.

**SELF-ASSESSMENT EXERCISE**

List two (2) characteristics of a good hypothesis?

## 3.3 Type I and Type II Errors

No hypothesis test is 100% certain since this test is based on probabilities. There is always a chance of making an incorrect conclusion. This is so since the decision to reject or accept the null hypothesis depends on the extent to which the computed value of the t statistic deviates from the expected value and bearing in mind there is the possibility of sampling and computation errors, one may make wrong decision. When a hypothesis is tested, it is possible to make two types of erroneous decisions. These two types of errors are referred to as type I and type II errors. The risks of these two errors are inversely related and determined by the level of significance and the power for the test.

**Type I Error**
Type one error is committed when the null hypothesis is true and one is led to reject it. The probability of making a type I error is denoted by the symbol "α", which is the level of significance you set for your hypothesis test.

**Type II Error**
Type one error is committed when the null hypothesis is false and one is led to accept it. The probability of making a type II error is denoted by the symbol "β", which is the level of significance you set for your hypothesis test.

Diagrammatically, it can be represented as
**Table 2.1: Definition of Type I and Type II Errors**

| | Reject | Fail to Reject |
|---|---|---|
| True Null Hypothesis | Type I Error (α) | No Error |
| False Null Hypothesis | No Error | Type II Error (β) |

**Self-Assessment Exercise 3**

Explain what you understand by type II error?

## 3.4 Level of Significance

The level of significance is defined as the probability of rejecting a null hypothesis by the test when it is really true. In testing a hypothesis, the maximum probability with which we would be willing to risk a Type I error is referred to as level of significance. Since in social science it is

not possible to eliminate sampling errors, it is thus left for the researcher to decide the margin of error being assigned to such sampling fluctuations in a study. It is conventional to fix the margin for error at 5%, 1% or 0.1%. this means the probability of rejecting the null hypothesis when it is in fact true is fixed at 0.05, 0.01 or 0.001 respectively. This probability often denoted by α is generally specified before any samples are drawn so that the results obtained will not influence one's choice. In practice, a significance level of 0.05 or 0.01 is customary, though other values can be used.

If say 0.05 (5%) is chosen as a significance level in designing a decision rule, then there are about 5 chances in 100 that the hypothesis will be rejected when it should be accepted; that is there is 95% confidence that the right decision was made. In such a case, it is said that the hypothesis has been rejected at 0.05 significance level. This means that the hypothesis has a 0.05 probability of being wrong. It is important to note that the lower the level of significance set, the higher the chances of rejecting a true null hypothesis. Thus there are far more chances of making type I error at 0.5 than at 0.05. Conversely, increasing the level of significance from 0.05 to 0.01 or 0.001 reduces the chances of making type I error, although it increases the chances of making a type II error.

**SELF-ASSESSMENT EXERCISE**

In determining the level of significance, what margin for error is acceptable?

## 3.5 The Critical Region

The critical region, also known as the region of rejection, is a set of values for a test statistic for which the null hypothesis is rejected in favour of the alternative hypothesis. This means that if the observed test statistic is in the critical region then we reject the null hypothesis and accept the alternative hypothesis. In any normal distribution, the regular events of common occurrence concentrate around the central values while the not so common events of rare occurrence are found at the extremes of the distribution. These unlikely events are located at the tail end of the curve. That area where these extreme values are located is known as critical region. This is so because on values of the test statistic that lie in this region are said to be statistically significant leading to a rejection of the null hypothesis.

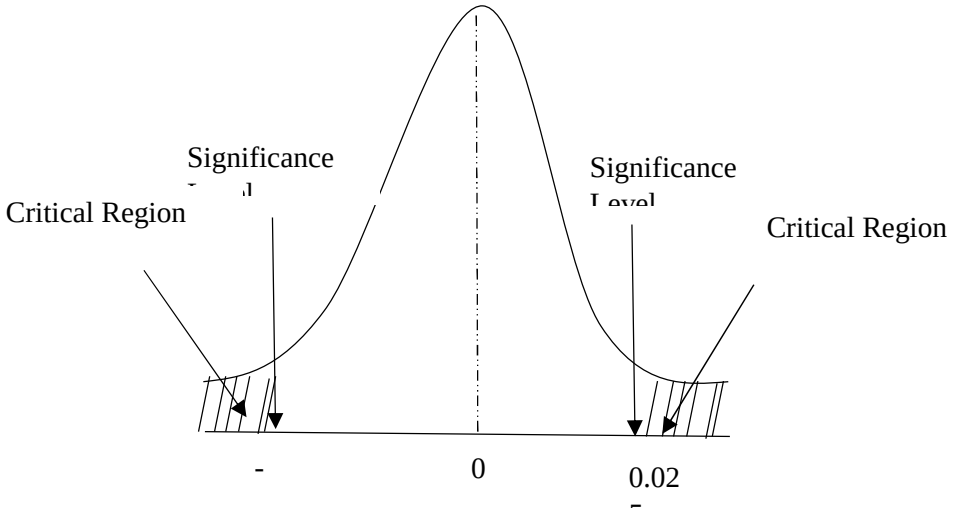



***Fig.2.1: The Normal Curve Showing the Critical Region***

**SELF-ASSESSMENT EXERCISE**

What is a critical region?

## 3.6 One-Tailed and Two-Tailed Tests

A one-tailed test is a statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. It is used when the research hypothesis is directional meaning that the hypothesis predicts or specifies the direction in which the variation will occur and one can test for effects in only one direction. When a one-tailed test is performed, the entire significance level percentage goes into the extreme end of one tail of the distribution. Since it is directional, the hypothesis will be in the form X is greater than Y or X is less than Y. when X is greater than Y, the critical region is located at the upper tail of the distribution which is positive, and when X is less than Y, the critical region is located at the lower tail of the distribution which is negative. If the sample being tested falls into the one-sided critical area, the null hypothesis is rejected and the alternative hypothesis will be accepted. For example, if a one tailed test is carried out on a hypothesis, a critical region covering the percentage of the level of significance say 0.01 or 0.05 is carved out at the upper tail end of the distribution as shown below

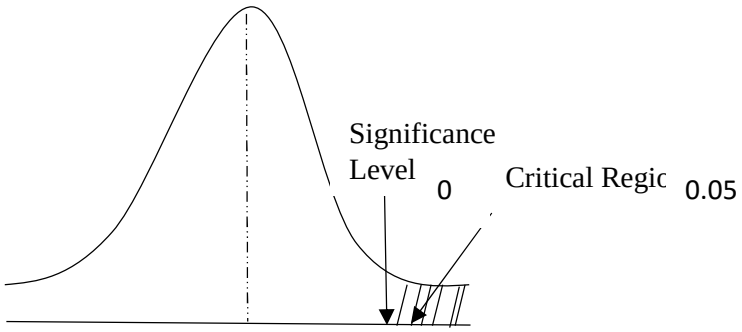



***Fig.2.2:One Tailed Test***

**Two Tailed Test**

Two-tailed test is also known as two-sided tests is a non-directional test in which the region of rejection or critical region for a hypothesis test is on both the ends of the normal distribution. This is so because one can test for effects in both directions. When you perform a two-tailed test, you split the significance level percentage between both tails of the distribution. This simply asserts that X differs from or is not equal to Y; meaning that the difference may be negative or positive.

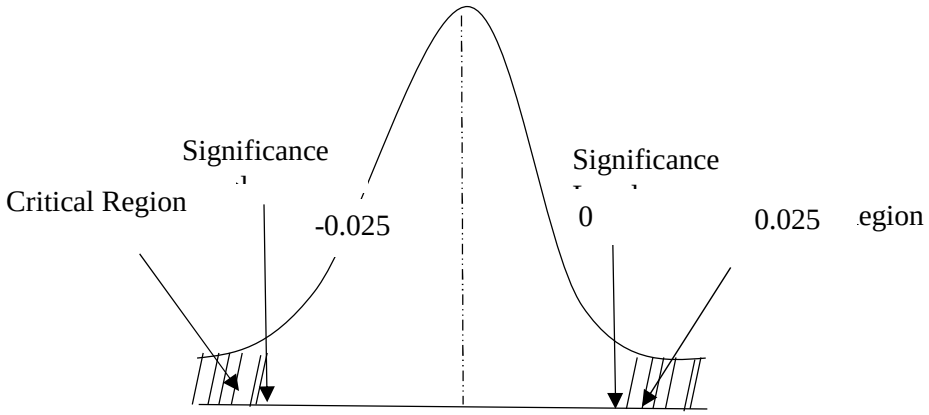



***Fig.2.3: Two Tailed Test***

**COMPARISON OF ONE-TAILED AND TWO-TAILED TESTS**

**Table 2.2: One-tailed and Two-tailed Tests Comparison**

| BASIS OF COMPARISON | ONE-TAILED TEST | TWO-TAILED TEST |
|---|---|---|
| Meaning | A statistical hypothesis test in which alternative hypothesis has only one end, is known as one tailed test | A significance test in which alternative hypothesis has two ends, is called two-tailed test |
| Hypothesis | Directional | Non-directional |
| Region of rejection | Either left or right | Both left and right |
| Relationship | If there is a relationship between variables in a single direction. | If there is a relationship between variables in either direction. |
| Result | Greater or less than certain value. | Greater or less than certain range of values. |
| Sign in alternative hypothesis | > or < | ≠ |

**SELF-ASSESSMENT EXERCISE**

Differentiate between a one-tailed and two-tailed test.

## 3.7 The Power of a Test

The power of a statistical test is the probability of rejecting a false null hypothesis. Power is the probability of making a correct decision. It is defined as 1 less the probability of type II error (1- β). The greater the ability of a test to eliminate false null hypothesis, the more powerful it is.

## 3.8 Degrees of Freedom (DF)

Degrees of freedom are the number of values that are free to vary as one estimates the parameters. It is the number of independent ways by which a system can move, without violating any constraint imposed on it. It is therefore, the number of independent values that a statistical analysis can estimate. The degrees of freedom can be calculated to help ensure the statistical validity of chi-square tests, t-tests and even the more advanced f-tests. These tests are commonly used to compare observed data with data that would be expected to be obtained according to a specific hypothesis and would be discussed. Degrees of freedom contribute to the validity of an outcome as it can identify how many

values in the final calculation are allowed to vary. These calculations are dependent upon the sample size, or observations, and the parameters to be estimated,
Mathematically, degrees of freedom is given as;
df = N-1,
Where N is the number of values in the data set (sample size).
For instance, if we have a data set (X) of 4, with the following values, 18, 35, 22, 17
The mean of the data set is (18+35+22+17)/4 = 23
Also, N=4.
This data set has a mean, or average of 23
Using the formula, the degrees of freedom (df) = N-1:
Therefore, df = 4-1 = 3
This indicates that, in this data set, three numbers have the freedom to vary as long as the mean remains 23.

**SELF-ASSESSMENT EXERCISE**

What is a degree of freedom in hypothesis testing?

## 3.9 Hypothesis Tests

There are several number of statistics used in testing the significance of the difference between proportions and between means. Three tests that are frequently employed by researchers include Chi Square test, the t test and the Z test. Here, we will be discussing the Chi Square and t test

### 3.9.1 Chi Square Test ($X^2$)

Chi-Square test is an inferential statistical test that measures how expectations compare to actual observed data. It is a test that assumes that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable. It thus tries to establish a relationship between given categorical variable in this test. The data used in calculating a chi square must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample. The test derives its name from the Greek capital letter Chi (X) pronounced “ki”. Chi square is frequently utilised in hypothesis testing because of its relationship to the normal distribution and it is a member of the class of likelihood ratio tests which have several desirable properties. However, the Chi-square test can only be used on the numbers. They can’t be used for percentages, proportions, means or similar statistical value.

On the whole, the rationale behind the Chi square rests in determining what a set of data would look like if the null hypothesis were to hold.

Also, for Chi Square not to give misleading results, data employed in the study must be in nominal form, sample should be randomly chosen, sample should consist of independent cases and no cell frequency should be less than five (5). The formula for calculating Chi Square is given as

$$X^2 = \sum \frac{(O-E)^2}{E}$$

or

$$X^2 = \sum (\frac{O^2}{E}) - N$$

Where
O = The observed frequencies
E = The expected frequencies

**Degree of Freedom in Chi Square Distribution**

The degrees of freedom in Chi Square distribution is equal to the number of standard normal deviates being summed. It is a measurement of the number of values in the statistic that are free to vary without influencing the result of the statistic. Since a statistical test is based on precise estimates and pieces of vital information, statisticians use these estimates to create statistical formulas that calculate the final result of their statistical analysis. The information used in analysis may vary, but there must always be at least one fixed category of information; the rest of the categories are degrees of freedom. This is important because statistics is based on hypotheses that can be hard to accurately compute. Degrees of freedom is therefore important in the Chi-Square test because the observed results often differ significantly from the expected results, and these degrees of freedom are needed to test different hypothetical situations.

Degree of freedom is calculated by using the following formula;

DF = (r-1)(c-1)

Where

DF = Degree of freedom
r = number of rows
c = number of columns

**Decision Rule**

The decision rule for Chi Square states that if the observed chi-square test statistic is greater than the critical value, the null hypothesis is rejected.

**Calculation of Chi Square Test**

**Example**

A survey was conducted on 145 industrial workers on whether tax incentives have impacted industrialisation in Nigeria. The question asked by the researcher was “Do you agree that tax incentive has impacted on economic growth?” Expected response was retrieved using

the ordinal scale ranking to evaluate relative feedback. The expected feedback was categorised into strongly agree, agree, undecided, disagree, strongly disagree. The table below shows the feedback based on responses

**Table 2.3: Respondent Feedback**

| Feedback | Respondents |
|---|---|
| Strongly agree | 19 |
| Agree | 48 |
| Undecided | 30 |
| Disagree | 26 |
| Strongly Disagree | 22 |
| Total | 145 |

In calculating the Chi Square, the following steps are necessary

**Step 1: State the Hypothesis**
Here we need to start by establishing a null hypothesis and an alternative hypothesis as given below.

**Null Hypothesis**
$H_0$: Tax incentives have no significant impact on industrialisation in Nigeria.

**Alternative Hypothesis**
$H_1$: Tax incentives have significant impact on industrialisation in Nigeria.

**Step 2: The sample analysis is carried out**
Here we will analyse the given sample data to compute

- Observed frequency
- Expected frequency
- Degree of freedom
- Calculate Chi-Square test static value

The observed frequency (O) are counts made from experimental data. In table 2.3 above, the observed frequencies are the number of respondents based on the ordinal scale ranked feedback. Thus, observed frequencies is stated in table 2.4 below.

**Table 2.4: Observed Frequencies**

| O |
|---|
| 19 |
| 48 |
| 30 |
| 26 |
| 22 |

**Expected frequency**

In the case of this example, the expected frequency is gotten by calculating the average of the observed frequency. Thus E $=\frac{19+48+30+26+22}{5} = 29$. The expected frequency is therefore 29.

Since we have gotten the observe and expected frequencies as shown in the table below, next is to calculate the degree of freedom (DF)

**Table 2.5: Observed and Expected Frequencies**

| O | E |
|---|---|
| 19 | 29 |
| 48 | 29 |
| 30 | 29 |
| 26 | 29 |
| 22 | 29 |

**Degree of Freedom**

The formula for DF = (r-1)(c-1)
Where r is the number of rows 2
C is the number of columns 5
Therefore, DF = (2-1) (5-1)
DF = 1 X 4 = 4
DF = 4
Next we calculate the Chi Square value. This is shown in the table below

**Table 2.6: Chi Square Calculation**

| | O | E | O – E | $(O-E)^2$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|---|---|
| | 19 | 29 | -10 | 100 | 3.45 |
| | 48 | 29 | 19 | 361 | 12.44 |
| | 30 | 29 | 1 | 1 | 0.034 |
| | 26 | 29 | -3 | 9 | 0.31 |
| | 22 | 29 | -7 | 49 | 1.69 |
| ∑ | | | | | 17.924 |

From the table above, we now know that Chi Square calculated ($X^2_{cal}$) = 17.92

Next we get the value of the tabulated Chi Square ($X^2_{tab}$). This is gotten using the Chi Square table.

($X^2_{tab}$) at 5% level of significance, where d.f = 4 and at 5% level of significance = 9.49
i.e.$X^2_{tab}$ = 9.49
From our finding, Chi Square calculated ($X^2_{cal}$) at 17.92 is greater than Chi Square from table ($X^2_{tab}$) at 9.49, we make our decision based on the finding.

**Decision rule**
Since Chi Square calculated is greater than Chi Square from table, we reject the null hypothesis.

**SELF-ASSESSMENT EXERCISE**

i. List three (3) test employed in testing the significance of a hypothesis
ii. Five (5) stores A, B, C, D and E compete for customer. The manager of store A hires a consultant to determine if the percentage of shoppers who prefer each of the five stores is the same. A survey of 1100 randomly selected shoppers is conducted, and the results about which one of the customers prefer is below. Is there enough evidence using a significance level $\alpha = 0.05$ to conclude that the proportions are really the same?

| Store | A | B | C | D | E |
|---|---|---|---|---|---|
| Customer | 262 | 234 | 204 | 190 | 210 |

## 3.9.2 T -Distribution

The T distribution, also known as the Student's t- test, is a type of probability distribution that estimates the population parameters when the sample size is small and the population standard deviation is unknown. T distributions have a greater chance for extreme values than normal distributions, hence the fatter tails.

**Properties of T- Distribution**
1. Like standard normal distribution the shape of the student distribution is also bell-shaped and symmetrical with mean zero.
2. The student distribution ranges from $-\infty$ to $\infty$ ($\infty$ means infinity).
3. The shape of the t-distribution changes with the change in the degrees of freedom.

4. The variance is always greater than one and can be defined only when the degrees of freedom greater or equal to 3.
5. It is less peaked at the center and higher in tails, thus it assumes platykurtic shape.
6. The t-distribution has a greater dispersion than the standard normal distribution. And as the sample size 'n' increases, it assumes the normal distribution. Here the sample size is said to be large when $n \geq 30$.

**Assumptions of t-Test**

1. The first assumption is concerned with the scale of measurement. Here assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale.
2. The second assumption is regarding simple random sample. The Assumption is that the data is collected from a representative, randomly selected portion of the total population.
3. The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve.
4. The fourth assumption is that that a reasonably large sample size is used for the test. Larger sample size means the distribution of results should approach a normal bell-shaped curve.
5. The final assumption is the homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.

**Types of t-Test**

Types of t test include;

- **One Sample t-Test**

  One sample t-test is a statistical procedure used to determine whether a sample of observations could have been generated by a process with a specific mean. To test this hypothesis, you could collect a sample of laptop computers from the assembly line, measure their weights, and compare the sample with a value of five using a one-sample *t*-test.

  The formula for the one-sample t-test for a population mean is the same as the z-test, except that the t-test substitutes the sample standard deviation *s* for the population standard deviation σ and takes critical values from the *t*-distribution instead of the *z*-distribution. The t-distribution is particularly useful for tests with small samples ($n < 30$).

  The purpose of the one sample *t*-test is to determine if the null hypothesis should be rejected, given the sample data. In a one

sample t test, there are two kinds of hypotheses; the null hypothesis and the alternative hypothesis. The alternative hypothesis assumes that some differences exist between the true mean (μ) while the null hypothesis assumes that there is no difference.

Mathematically, one sample t test is calculated using the formula

$t = \frac{\bar{x}-\mu_o}{s/\sqrt{n}}$ df = n-1

- **Two Sample t-Test**

  A two-sample t test is used to determine if two population means are equal. In independent sample t test, two samples that are not at all related to each other are tested. The main idea behind two independent sample t tests is to draw out a statistical inference about the comparison of two independent samples of data.

  In the case of a two sample test, the data may either be paired or not paired. Paired t test means that there is a one-to-one correspondence between the values in the two samples. That is, if $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are the two samples, then $X_i$ corresponds to $Y_i$. For unpaired samples, the sample sizes for the two samples may or may not be equal.

  The sampling distribution for a two sample t test is given as

  $t = \frac{\bar{X}-\bar{X}_2}{s_{x_1x_2}*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ df = $n_1 + n_2 - 2$

  For the paired t tests the sampling distribution is given as

  $t = \frac{\bar{X}_D-\mu_o}{s_D/\sqrt{n}}$ df = n-1

**Examples**

A teacher wants to know if his statistics class has a good grip of mathematics. Six students are chosen at random from the class for proficiency test. The professor wants the class to be able to score above 70 on the test. The students' scores include of 62, 92, 75, 68, 83, and 95. Can the teacher have 90 percent confidence that the mean score for the class on the test would be above 70?

**Solution:**

Null hypothesis: $H_0$: μ = 70

Alternative hypothesis: $H_1$: μ > 70

First, compute the sample mean and standard deviation:

The mean for the scores is $\frac{62+92+75+68+83+95}{6}$ =79.17

Using the standard deviation formula in module 2, the value for the standard deviation = 13.17
Next we compute the *t*-value
Remember that
$\bar{X} = 79.17$, $\mu_o = 70$, $s = 13.17$, n = 6
Therefore $t = \frac{79.17-70}{13.17/\sqrt{6}}$

$$t = \frac{9.17}{5.38}$$

$$t = 1.71$$

Next we test the hypothesis
To test the hypothesis, the computed *t*-value of 1.71 will be compared to the critical value in the *t*-table.
Remember that the teacher need a 90 percent confidence
90 percent confidence level = 0.10.
The extreme values in one rather than two directions, thus a one-tailed test, and you do not divide the alpha level by 2.
Degrees of freedom (DF) = n-1
6 – 1 = 5.
Using the t-table for $t_{0.10}$, the value of the t is 1.476.
Since the computed *t*-value of 1.71 is larger than the critical value in the table 1.476, we reject the null hypothesis and conclude that the professor has evidence that the class mean on the mathematic test would be at least 70.

**SELF-ASSESSMENT EXERCISE**

Discuss types of t test.

## 4.0 CONCLUSION

From our discussion on this unit, you have learnt the following:

- Hypothesis and its types, importance and characteristics
- Type I and Type II errors and Level of Significance
- One-tailed and two-tailed tests and its comparison
- Hypothesis testing using Chi Square and t test.

## 5.0 SUMMARY

It is expected that at this stage, students should be able to differentiate between a substantive and a null hypothesis, understand the importance and characteristics of a hypothesis and thus be able to build hypothesis. Also, students are expected to be able to identify type I and type II errors

among other things. Finally, they are to be able to test hypothesis using either the chi square or the t test.

## 6.0 TUTOR-MARKED ASSIGNMENT

What is the difference between one-tailed and two-tailed test?

## 7.0 REFERENCES/FURTHER READING

Damodar, N. G., Dawn, C. P., & Sangetha, G. (2012). *Basic Econometrics.* New Delhi: Tata McGraw Hill Education Private Ltd.

Obikeze D. S (1986). *Introductory Statistics for the Social Sciences.* Enugu: Fourth Dimension Publishing Co. Ltd.

# UNIT 3 CORRELATION

## CONTENTS



## 1.0 INTRODUCTION

In this unit, we shall study correlation. Correlation deals with measuring the strength and direction of the linear relationship between two variables usually a dependent variable and an independent variable(s) in a model. Types of correlations (Spearman rank and Pearson) and the coefficient of determination denoted by $R^2$ will be explained and studies

## 2.0 OBJECTIVES

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

- define correlation
- explain types of correlation
- calculate correlation using the Spearman rank and Pearson correlations
- identify and explain the properties of Pearson correlation.

## 3.0 MAIN CONTENT

## 3.1 Correlation

Correlation is an analysis that measures the strength of association and the direction of relationship between two variables. It is used to test relationships between quantitative variables. In other words, it is a measure of how things are related and it is useful because if you can find

out what relationship variables have with predictions about future behavior possible. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. A value of +1 or -1 indicates a perfect degree of association between the two variables. However, when the value is zero (0), then, no relationship exists. Thus, it becomes clear that as the correlation coefficient value goes towards 0, the relationship between the two variables becomes weaker. The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a – sign indicates a negative relationship.

**SELF-ASSESSMENT EXERCISE**

In statistics, what do you understand by the term correlation?

## 3.2 Types of Correlations

Generally, we have four types of correlations:

- Pearson correlation
- Kendall rank correlation
- Spearman correlation
- Point-Biserial correlation.

Of these types of correlation, the Pearson Correlation and Spearman Correlation are the most widely used and will be discussed here.

### 3.2.1 Pearson Correlation

The Pearson product-moment correlation is a correlation statistic used to measure the degree of the relationship between linearly related variables. It was developed by Karl Pearson and it is used to measure the degree of linear relationship between the two variables. It is the covariance of the two variables divided by the product of their standard.

The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample. Since we will be dealing with samples, we will use "r" to represent Pearson's correlation. Pearson's r can range from -1 to 1 and may be either positive, negative or have no linear relationship between variables.

### 3.2.2 Properties of Pearson Correlation

a) Relationship ranges between -1 to 1.
b) Pearson's correlation is symmetric. This means that the correlation of X with Y is the same as the correlation of Y with X.

c) Pearson's r is not affected by linear transformations. This means that multiplying a variable by a constant and/or adding a constant does not change the correlation of that variable with other variables.

d) Pearson's r is independent of the unit of measurement

The formula for calculating Pearson correlation is given as

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Where $r_{xy}$= Pearson r correlation coefficient between x and y
$n$ = number of observations
$x_i$= value of x (for $i^{th}$ observation)
$y_i$ = value of y (for $i^{th}$ observation)

An alternative computational formula that avoids the step of computing deviation scores is

$$r_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{(\sum x^2 - \frac{\sum x}{N})}\sqrt{(\sum y^2 - \frac{(\sum y)^2}{N})}}$$

Example: Given the values of X and Y in the table below, calculate the Pearson correlation

| | | | | | |
|---|---|---|---|---|---|
| X | 2 | 4 | 6 | 6 | 7 |
| Y | 5 | 7 | 11 | 13 | 14 |

$\bar{X} = 5$

$\bar{Y} = 10$

Remember that

$x = X - \bar{X}$

$y = Y - \bar{Y}$

$xy = (X - \bar{X})(Y - \bar{Y})$

| | X | Y | $X$ | $Y$ | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|---|---|---|
| | 2 | 5 | -3 | -5 | 15 | 9 | 25 |
| | 4 | 7 | -1 | -3 | 3 | 1 | 9 |
| | 6 | 11 | 1 | 1 | 1 | 1 | 1 |
| | 6 | 13 | 1 | 3 | 3 | 1 | 9 |
| | 7 | 14 | 2 | 4 | 8 | 4 | 16 |
| $\sum$ | 25 | 50 | 0 | 0 | 30 | 16 | 60 |

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

$$r_{xy} = \frac{30}{\sqrt{16X60}}$$

$$r_{xy} = \frac{30}{\sqrt{960}}$$

$$r_{xy} = \frac{30}{30.98}$$

$= 0.97$

**Using the alternative formula**

$$r_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{(\sum x^2 - \frac{\sum x}{N})}\sqrt{(\sum y^2 - \frac{(\sum y)^2}{N})}}$$

$$r_{xy} = \frac{30 - \frac{0(0)}{5}}{\sqrt{(16 - \frac{0}{5})}\sqrt{(60 - \frac{(0)^2}{5})}}$$

$$r_{xy} = \frac{30 - 0}{\sqrt{(16 - 0)}\sqrt{(60 - 0)}}$$

$$r_{xy} = \frac{30 - 0}{\sqrt{16}\sqrt{60}}$$

$$r_{xy} = \frac{30}{4\ X\ 7.75}$$

$$r_{xy} = \frac{30}{31}$$

$$r_{xy} = 0.97$$

### 3.2.3 Spearman Rank Correlation Coefficient

The Spearman rank correlation is a non-parametric test used to measure the degree of association (relationship) between two variables. The Spearman rank correlation test does not carry any assumptions about the distribution of the data. It rather uses ranks and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal. It is denoted either with the Greek letter *rho* (ρ), or *r*. It measures the strength of association between two variables.

The formula used to calculate the Spearman rank correlation is stated as:

$$r = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

**Example:** Given the table below, compute the Spearman rank correlation coefficient

| Convenience Store | Distance from Jabi (m) | Rank Distance | Price of 50cl bottle (N) | Rank price |
|---|---|---|---|---|
| 1 | 50 | 10 | 1.80 | 2 |
| 2 | 175 | 9 | 1.20 | 3.5 |

| | | | | |
|---|---|---|---|---|
| 3 | 270 | 8 | 2.00 | 1 |
| 4 | 375 | 7 | 1.00 | 6 |
| 5 | 425 | 6 | 1.00 | 6 |
| 6 | 580 | 5 | 1.20 | 3.5 |
| 7 | 710 | 4 | 0.80 | 9 |
| 8 | 790 | 3 | 0.60 | 10 |
| 9 | 890 | 2 | 1.00 | 6 |
| 10 | 980 | 1 | 0.85 | 8 |
| | | | | |

First, we find the value of the difference between the ranks and then we square the difference ($d^2$). This is explained by the table below:

| Convenience Store | Distance from Jabi (m) | Rank Distance $R_1$ | Price of 50cl bottle (N) | Rank price $R_2$ | Difference between ranks (d) $(R_1 - R_2)$ | $d^2$ $(R_1 - R_2)^2$ |
|---|---|---|---|---|---|---|
| 1 | 50 | 10 | 1.80 | 2 | 8 | 64 |
| 2 | 175 | 9 | 1.20 | 3.5 | 5.5 | 30.25 |
| 3 | 270 | 8 | 2.00 | 1 | 7 | 49 |
| 4 | 375 | 7 | 1.00 | 6 | 1 | 1 |
| 5 | 425 | 6 | 1.00 | 6 | 0 | 0 |
| 6 | 580 | 5 | 1.20 | 3.5 | 1.5 | 2.25 |
| 7 | 710 | 4 | 0.80 | 9 | -5 | 25 |
| 8 | 790 | 3 | 0.60 | 10 | -7 | 49 |
| 9 | 890 | 2 | 1.00 | 6 | -4 | 16 |
| 10 | 980 | 1 | 0.85 | 8 | -7 | 49 |
| | | | | | $\sum$ | 281.5 |

$$r \approx 1 - \frac{6X281.5}{10(10^2 - 1)}$$

$$r = 1 - \frac{1689}{10(100 - 1)}$$

$$r = 1 - \frac{1689}{10(99 - 1)}$$

$$r \approx 1 - \frac{1689}{10(98)}$$

$$r = 1 - \frac{1689}{980}$$

$$r = 1 - 1.72$$

$$r = -0.72$$

**SELF-ASSESSMENT EXERCISE**

i. What are the properties of Pearson correlation?
ii. What formula is used to calculate the Spearman rank correlation

## 3.2.4 Coefficient of Determination ($R^2$)

The coefficient of determination is a measure used in statistical analysis that assesses how well a model explains and predicts future outcomes. It is a measure of goodness of linear function. The coefficient of determination, also commonly known as 'R squared" ($R^2$) is used as a guideline to measure the accuracy of the model. It gives you the percentage variation in the dependent variable (Y) that is explained by the independent variable(s) X. It is important to note that coefficient of determination ranges from 0 to 1 (i.e. 0% to 100% of the variation in Y can be explained by X) and the larger the coefficient of determination, the better the fit.

$$R^2 = 1 - \frac{First\ Sum\ of\ Errors}{Second\ Sum\ of\ Errors}$$

It is also calculated using the formula

$$R^2 = 1 - \frac{S^2y.x}{S^2y}$$

Where $S^2y.x = \frac{(Y-Y^1)^2}{N}$ and $S^2y = \frac{(Y-\bar{Y})}{N}$

**SELF-ASSESSMENT EXERCISE**

What range does the coefficient of determination fall into?

## 4.0 CONCLUSION

Correlation was explained in this unit. You were exposed to types of correlation and examples on how to calculate correlation using different formulae. This was done to show the relationship between variables being tested.

## 5.0 SUMMARY

In this unit, we were able to discuss in detail correlation and the relationship between the independent variable and the dependent variable was shown. Four types of correlation were listed and two of them were discussed extensively and calculations were done using their different formulae.

## 6.0 TUTOR MARKED ASSIGNMENT

1. What is correlation?
2. Given the values of X and Y in the table below, calculate the Pearson correlation.

| | | | | | |
|---|---|---|---|---|---|
| X | 3 | 5 | 8 | 8 | 9 |
| Y | 6 | 8 | 10 | 12 | 16 |

## 7.0 REFERENCES/ FURTHER READING

Bobko, P. (2001). *Correlation and Regression: Applications for Industrial Organizational Psychology and Management* (2nd ed.). Thousand Oaks, CA: Sage Publications

Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and Nonparametric Measures.* Thousand Oaks, CA: Sage Publications.

Dominick, S. & Derrick, R. (2011). *Statistics and Econometrics (Schaum Outlines)*. NewYork: McGraw-Hill Company,

Obikeze D. S (1986). *Introductory Statistics for the Social Sciences.* Enugu Fourth Dimension Publishing Co. Ltd.

Okoro, E. (2002). *Quantitative Techniques in Urban Analysis.* Ibadan: Kraft Books Ltd.

# UNIT 4 REGRESSION ANALYSIS

## CONTENTS



## 1.0 INTRODUCTION

Regression analysis is an analytical tool employed in statistical research to explain the functional relationships among variables. It attempts to explain the effect of the dependence of one variable on another. This unit is essential as it equips students with the ability to know how to use and apply regression analysis in the practical research. For a student to be successful in this course, he/she is required to have thorough knowledge of simple regression model and hypothesis testing, among others.

## 2.0 OBJECTIVES

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

- Carry out a regression analysis
- state parameter estimates involved
- explain how to estimate the parameters such as $b_o$, $b_1$, $b_2$, …bn
- discuss the uses and types of regression analysis.

## 3.0 MAIN CONTENT

### 3.1 Meaning of Regression Analysis

Regression analysis is a statistical process used to examine/determine the relationship between two or more variables (one dependent and one or more independent variable(s)). The variable that is to be estimated is

referred to as the dependent variable and is denoted by the symbol "Y" while the variable(s) that used in the estimation is (are) called the independent variable(s) or predictor(s) or predetermined variable(s). The value of the independent variable is known and could be controlled by the investigator while the value of the former is determined by the independent variable(s). This implies that the dependent variable is a function of the independent variables, i.e. Y = f(X).

The term regression was first applied to statistics by the statistician Francis Galton in 1877 in his study on the relationship between the characteristics of parents and their children. He tried to describe the tendency of children to regress to the average level of their parents in a number of traits. In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable Y and one or more independent variables denoted by X. In the case of one independent variable, it is called a simple linear regression.

As earlier stated, a simple linear regression has only one dependent variable and one independent variable. When a regression analysis has more than one independent variable, it is called a multiple regression analysis, or variables (in the event of multiple regression). The word, linear means that the regression analysis is a straight line graph.

Mathematically, the simple linear regression model is stated thus:

$Y_i = \alpha_i + \beta_i X + \mu_i$----------(1)

It should be noted that in this model,

- α is a constant. It represents the point at which the regression line crosses the Y axis. It is also referred to as the intercept and can be positive, negative or zero. The positive value indicates that the regression line crosses the Y axis from above the X axis, the negative value indicates that the regression line crosses the Y axis from below the X axis, while it is zero, it indicates that the regression line passes through the origin.
- β is a notation used to represent the parameter.
- X is the independent variable
- Y is the dependent variable
- μ is the random or stochastic error term

Note: The error term accounts for the variability in Y that cannot be explained by the linear relationship between X and Y. It therefore represents all the other variables not clearly included in the linear regression model.

**SELF-ASSESSMENT EXERCISE**

Define the term, regression,

## 3.2 Uses of Regression Analysis

The major uses of regression analysis include;

1. To determine the strength of predictors: The regression might be used to identify the strength of the effect that the independent variable(s) have on the dependent variable. Typical questions are what is the strength of relationship between interest rate and investment in the economy?

2. To forecast an effect: Regression analysis can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables.

3. Trend forecasting: Thirdly, regression analysis is used to predict trends and future values. The regression analysis can be used to get point estimates- for example what will the inflation rate in the economy in 6 months' time?

**SELF-ASSESSMENT EXERCISE**

Enumerate the uses of regression analysis.

## 3.3 Types of Regression Analysis

The following are some of the types of regression analysis

**1. Linear Regression**

A linear regression refers to a regression model that is completely made up of linear variables. It is a technique used to model the relationship between an independent variable(s) as input and an output dependent variable using a linear model i.e. a straight line.

**2. Polynomial Regression**

Polynomial regression is a model that is used in handling non-linearly separable data. In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points. For a polynomial regression, the power of some independent variables is more than 1.

**3. Ridge Regression**

A standard linear or polynomial regression will fail in a case where there is high collinearity among the feature variables. The ridge regression becomes an option. Ridge Regression is a technique for analysing regression data that suffer from multi-collinearity. Collinearity is the existence of near-linear relationships among the independent variables.

**4. Lasso Regression**

Lasso regression analysis is a shrinkage and variable selection method for linear regression models. The goal of lasso regression is to obtain the subset of predictors that minimize prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.

Of these types of regression analysis, the linear regression is the most common and frequently used. Linear regression is generally classified into two types; simple linear regression and multiple linear regression.

Simple linear regression deals with a linear relationship between one dependent and one independent variable. Meanwhile multiple linear regression deals with a linear relationship between one dependent and two or more independent variables.

**SELF-ASSESSMENT EXERCISE**

List and discuss two types of regression analysis

## 3.4 Possible Regression Lines in Simple Linear Regression

**Positive Linear Relationship**

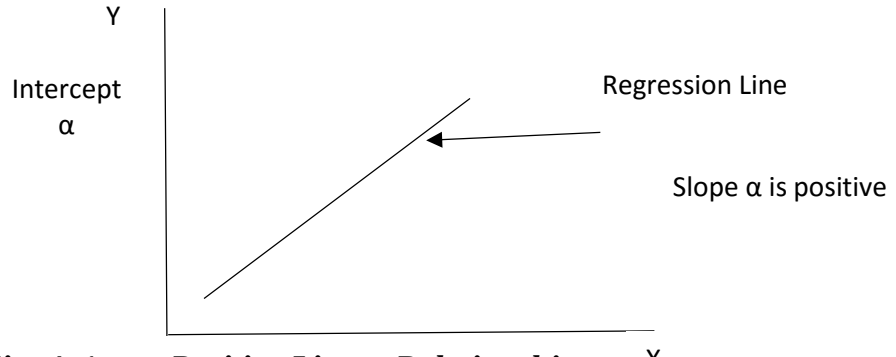



***Fig. 4. 1 : Positive Linear Relationship***

Here, the relationship between both variables is positive. i.e. as variable X is increasing, Y is decreasing. Thus, the slope is downwards.

Examples include increase in input leads to increase in output, increase in consumer spending and GDP etc.

**Negative Linear Relationship**

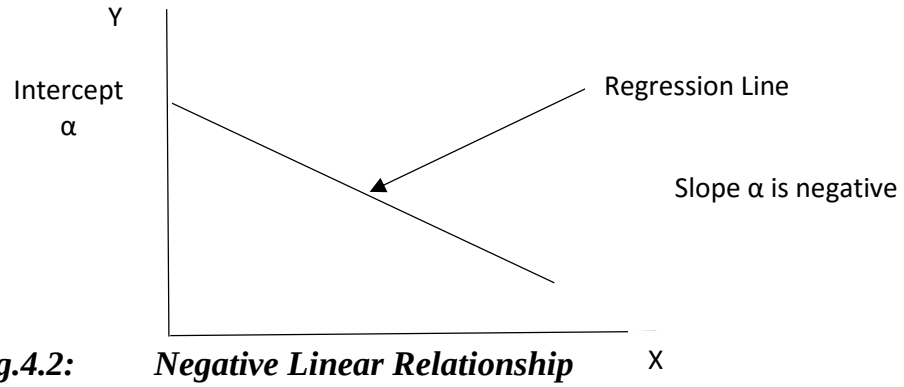



***Fig.4.2: Negative Linear Relationship***

Here, the relationship between both variables is negative. i.e. as variable X is increasing, Y is also increasing. Thus, the slope is upwards. Examples include; increase in lending rate leads to decrease in investment, increase in age of a dairy cow lead to decrease in milk she produces etc.

**No Linear Relationship**

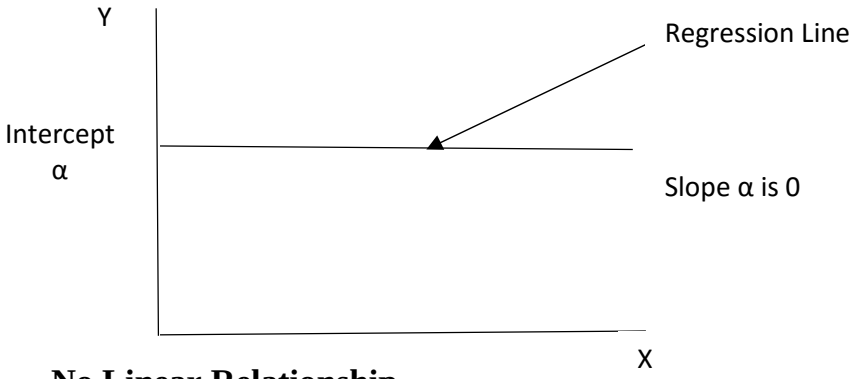



**Fig.4.3: No Linear Relationship**

Here, there is no relationship between both variables. Thus, the slope is horizontal to the origin. Examples include there is no relationship between tea drunk and intelligence, good governance and ethnic origin etc.

**SELF-ASSESSMENT EXERCISE**

List three (3) pairs of variables perceived to have a positive relationship.

## 3.5 Assumption of Linear Regression Analysis

The assumptions of linear regression include the following:

1. Linear relationship. This means that the regression of Y on X takes the form of a straight line
2. Bivariate normality. This means that the two variables X and Y form a normal distribution. i.e. the two variables are normally distributed about each other.
3. Homoscedasticity. This implies that the variances of the Y distributions are the same for each value of X.
4. No or little multicollinearity: absence of linear relationship among variables
5. No auto-correlation

In discussing the simple linear regression analysis, we shall apply two common methods: the graphical method and the least square method.

## 3.6 Least Square Method of Regression Analysis

The least squares method is a procedure for using sample data to find the estimated regression equation.

**Illustration**
To illustrate the least squares method, suppose data were collected as shown below

**Table 4.1: Table showing two variables X and Y**

| | | | | | | |
|---|---|---|---|---|---|---|
| X | 8 | 11 | 13 | 14 | 15 | 17 |
| Y | 30 | 25 | 22 | 21 | 20 | 18 |

From the above table and based on economic theory, quantity demanded of a commodity is a function of its price. Thus price is the dependent variable (X) while quantity demanded is the independent variable (Y). The values of X and Y for the sample are summarized below

**Table 4.2: Table showing two variables X and Y**

| Y | X |
|---|---|
| 30 | 8 |
| 25 | 11 |
| 23 | 13 |
| 22 | 14 |
| 20 | 15 |
| 18 | 17 |

A scatter diagram (scattergram) of the data in table above is plotted with quantity demanded (X) on the horizontal axis and price (Y) on the vertical axis.

Note that scatter diagrams for regression analysis are constructed with the independent variable X on the horizontal axis and the dependent variable Y on the vertical axis. The scatter diagram enables us to observe the data graphically and to draw preliminary conclusions about the possible relationship between the variables. The scattergram is shown below

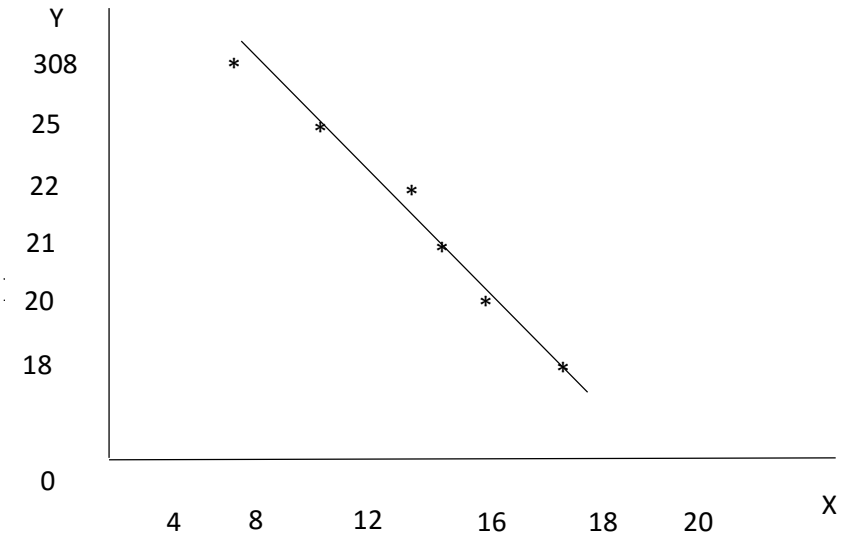



***Fig. 4.4: Scattergram showing relationship between X and Y***

From the graph, the relationship between the quantity demanded and price of a commodity appears to be approximately a straight line; indeed, a positive linear relationship is indicated between X and Y. We therefore choose the simple linear regression model to represent the relationship between quantity demanded and price. Given that choice, our next task is to use the sample data to determine the values of the intercept α and the parameter β in the estimated simple linear regression equation.

The regression is given as $Y_i = \alpha_i + \beta_i X + \mu_i$
Where Y = price and X = quantity demanded.

The formula for calculating the values of α and β include;

$$\alpha = \frac{(\sum X^2)(\sum Y) - (\sum X)(\sum XY)}{n\sum X^2 - (X)^2}$$

$$\beta = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (X)^2}$$

α and β can also be obtained by the following formulae which are expressed in deviation of the variables from their mean

$$\alpha = \bar{Y} - \alpha\bar{X}$$

$$\beta = \frac{xy}{x^2}$$

Where $x = X - \bar{X}$

$$y = Y - \bar{Y}$$

$$\bar{X} = Mean\ of\ X$$

$$\bar{Y} = M \quad of\ Y$$

Calculating the example above, we begin by estimate the mean value of X and Y

$\bar{X} = \frac{8+11+13+14+15+17}{6} = \frac{78}{6} = 13$

$\bar{Y} = \frac{30+25+23+22+20+18}{6} = \frac{138}{6} = 23$

Remember that

$x = X - \bar{X}$

$$y = Y - \bar{Y}$$

$xy = (X - \bar{X})(Y - \bar{Y})$

**Table 3.9: Regression Table**

| | Y | X | *Y* | *x* | $x^2$ | $y^2$ | *xy* |
|---|---|---|---|---|---|---|---|
| | 30 | 8 | 7 | -5 | 25 | 49 | -35 |
| | 25 | 11 | 2 | -2 | 4 | 4 | -4 |
| | 23 | 13 | 0 | 0 | 0 | 0 | 0 |
| | 22 | 14 | -1 | 1 | 1 | 1 | -1 |
| | 20 | 15 | -3 | 2 | 4 | 9 | -6 |
| | 18 | 17 | -5 | 4 | 16 | 25 | -20 |
| ∑ | 138 | 78 | 0 | 0 | 50 | 88 | -66 |

Therefore

$\beta = \frac{xy}{x^2}$

$\beta = \frac{-66}{50}$

$\beta = -1.32$

$\alpha = \bar{Y} - \alpha\bar{X}$

$\alpha = 23 - (-1.32 \text{ x } 13)$

$\alpha = 23 + 17.16$

$\alpha = 40.16$

Thus, the estimated regression equation is $Y = 40.16 - 1.32X$. This is also known as the regression line.

The value of the estimated parameter β is negative, implying that as price of the commodity increases, quantity demanded falls.

Given that the least squares estimated regression equation adequately describes the relationship between X and Y, it would seem reasonable to use the estimated regression equation to predict the value of Y for a given value of X.

**SELF-ASSESSMENT EXERCISE**

The table below shows an observation collected in a regression study on two variables.

| | | | | | |
|---|---|---|---|---|---|
| X | 2 | 6 | 9 | 13 | 20 |
| Y | 7 | 18 | 23 | 26 | 23 |

a. Develop a scatter diagram for these data.
b. Show the regression equation

## 4.0 CONCLUSION

The unit successfully explained regression analysis for you to understand and apply. So, you have learnt:

- Simple linear regression
- Calculation of regression analysis using least square method

## 5.0 SUMMARY

In the unit, we have been able to carry out a regression of an independent variable on the dependent variable, understand parameter estimates and it's working.

## 6.0 TUTOR-MARKED ASSIGNMENT

1. What is a linear regression?
2. Explain the assumptions of a linear regression
3. The table below shows an observation collected in a regression study on two variables.

| | | | | | |
|---|---|---|---|---|---|
| X | 2 | 6 | 9 | 13 | 20 |
| Y | 7 | 18 | 23 | 26 | 23 |

a. Develop a scatter diagram for these data.
b. Develop the estimated regression equation for these data.
c. Use the estimated regression equation to predict the value of Y when X = 4.

## 7.0 REFERENCES/ FURTHER READING

Bobko, P. (2001). *Correlation and Regression: Applications for Industrial Organizational Psychology and Management* (2nd ed.). Thousand Oaks, CA: Sage Publications

Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and Nonparametric Measures.* Thousand Oaks, CA: Sage Publications.

Dominick, S. & Derrick, R. (2011). *Statistics and Econometrics (Schaum Outlines).* NewYork: McGraw-Hill Company,

Obikeze D. S (1986). *Introductory Statistics for the Social Sciences.* Enugu Fourth Dimension Publishing Co. Ltd.

Okoro, E. (2002). *Quantitative Techniques in Urban Analysis.* Ibadan: Kraft Books Ltd.