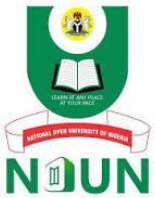
COURSE GUIDE

STT 205
STATISTICS FOR MANAGEMENT SCIENCES 1

Course Team Dr. Chigozie Kelechi Acha (Course Developer)-NOUN
Michael Okpara (Course Writer)-University of Agriculture, Umudike, Abia State
Prof. Peter A. Osannaiye (Content Editor)-University of Ilorin, Ilorin
Instructional Designer:
Learning Technologists:
Copy Editor



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

First Printed 2023, Printed 2024

ISBN: 978-978-786-034-2

CONTENTS

INTRODUCTION

Statistics for Management Sciences 1 (STT 205) is a core course in first semester and 3 credit units second year level course in the School of Management Sciences at the National Open University, Nigeria. STT 205 is on with Theories and Practicals of Statistics in general. The course consists of twenty-two units that involves basic concepts and principles of statistic and decision-making process, forms of data, summarizing data, graphical presentation of data, measures of both locations and dispersion, methods of data estimation, some elements of hypothesis testing, probabilities and statistical distributions of both discrete and continuous random variables. It also contains progressive statistical methods, introduction to parametric methods and test based on runs, and fundamentals of index numbers. It serves as a fundamental guide in statistics to students as it presents almost all major aspects of statistics in a simplified manner.

The course is designed to introduce students to statistics which is why it focuses on the fundamental principles of statistics with the intention of building a solid foundation. It requires that students study the course materials carefully, supplement the materials with other resources from statistics textbooks that treat the contents.

COURSE COMPETENCIES

The overall goal of STT 205 Statistics for Management Sciences I is to open the vista of the introductory aspects of statistics to students, where estimation, prediction, and decision making comes from analysis of properly collected data. It also, presents statistical data collection as an operation of statistical data processing aimed at gathering of statistical data and producing the input object data of a statistical survey. In addition, different sources and methods of data collection are included. Besides, it covers the descriptive handling of statistical data and looks at diagrammatic representation - diagrams, charts and graphs. Furthermore, it provides the process of deducing properties of underlying distributions by analysis of data, known as statistical inference. It also has in it measures of central tendency (location) without leaving the measures of dispersion (spread) which is a lack of uniformity in the sizes or quantities of the items of a group or series. Elementary probability (special probability distributions) and Statistical (probability) distributions are discussed. An x-ray of the basic estimates (point and interval) and principle of hypothesis testing details are considered. Other statistical methods covered include nonparametric statistical methods and index numbers.

The modes of presentation of the topics makes for good understanding of the techniques involved. In addition, exercises are provided at the end of the units to assist readers test their understanding of the topics covered.

COURSE OBJECTIVES

On completing this course, the student is expected to be able to:

1. Understand the usefulness and applicability of Statistics in his/her area of study;
2. Showcase adequate skills in problem solving involving various descriptive and inferential statistics and probability;
3. Adequately describe any given data using the measures of location, partition and dispersion;
4. Understand how to estimate a population parameter under different conditions
5. State and test a hypothesis and arrive at a statistical conclusion about populations, especially those from his/her area of study;
6. Develop the skill of describing real life situations in his/her area in probabilistic form
7. Identify and apply basic probability distributions governing different experimental research
8. Arrange and carry out statistical tests on data in a contingency table
9. Adequately handle regression and correlation problems
10. Develop basic statistical skills for handling Design and Analysis of Experiment data and Analysis of variance.

The course major aim is to give the students a detailed and simplified yet standard approach to the understanding of statistical techniques useful in presentation, description and analysis of statistical data. It will also expose the students to decision making techniques in statistical inference. It will as well help the students to become adequately conversant with statistical techniques required to successfully prosecute their academic programmes, especially skills needed to carry out their final year research project and those required in future research works and development.

WORKING THROUGH THIS COURSE

The course is divided into modules and units. The modules are derived from the course competencies and objectives. The competencies will guide you on the skills you will gain at the end of this course. So, as you work through the course, reflect on the competencies to ensure mastery. The units are components of the modules. Each unit is sub-divided into ii introduction, intended learning outcome(s), main content, self-

assessment exercise(s), conclusion, summary, and further readings. The introduction introduces you to the unit topic. The intended learning outcome(s) is the central point which help to measure your achievement or success in the course. Therefore, study the intended learning outcome(s) before going to the main content and at the end of the unit, revisit the intended learning outcome(s) to check if you have achieved the learning outcomes. Work through the unit again if you have not attained the stated learning outcomes. The main content is the body of knowledge in the unit. Self-assessment exercises are embedded in the content which helps you to evaluate your mastery of the competencies. The conclusion gives you the takeaway while the summary is a brief of the knowledge presented in the unit. The final part is the further readings. This takes you to where you can read more on the knowledge or topic presented in the unit.

You are required to read the study units, textbooks and other materials on the course.

Each unit contains self-assessment exercises called Student Assessment Exercises, (SAE). At some points in the course, you are required to write Tutor-Marked Assignments **(**TMA), Computer Based Test (CBT) and submit on NOUN TMA PORTAL for assessment. At the end of the course there is a final Examination. This course should take about 15 weeks to complete. Some listed components of the course, what you have to do and how you should allocate your time to each unit in order to complete the course successfully on time, are given below.

COURSE MATERIALS

Major components of the STT 205 are:
(1) Course Guide
(2) Study Units
(3) Textbooks
(4) Assignment File
(5) Presentation Schedule.

COURSE GUIDE

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

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

STUDY UNITS

Modules and Units

The first three units in Module 1 are on the nature of statistics and basic concepts in statistics. Units 4, 5, 6 and 7 in Module 2, constitute different presentations of statistical data. The subsequent three units in Module 3 are focused on computation of measures of central tendency and dispersion from samples and populations. Units 11 to 14 in Module 4 are on basic concepts and principles of probability (set theory and special probability distributions**)**. Furthermore, units 15, 16 and 17under Module 5 are on statistical distributions, Module 6 is on Estimation and Test of Hypothesis while the last two units, Module 7, teaches the progressive statistical methods.

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

Each unit contains a number of self-tests. In general, these self-tests question you on the material you have just covered or required you to apply it in some way and thereby help you to assess your progress and to reinforce your understating of the material. These exercises together with tutor-marked assignments will assist at achieving the stated learning objectives of the individual units and of the course. It is pertinent to note that the course is in seven modules and twenty-two study units as shown under study contents;

REFERENCES/FURTHER READING/WEB RESOURCES

Recommended Textbooks
It is advisable you have some of the following books;

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

Bhattacharyya, G. K., and R. A. Johnson, (1997). Statistical Concepts and Methods, John Wiley and Sons, New York.

Buglear, J. (2002). Stats Means Business: A Guide to Business Statistics, Butterworth-Heinemann, Oxford, ISBN 0-7506-5364-7.

Dodge, Y. (2008). The Concise Encyclopedia of Statistics. Springer. Everitt, B. S.; Skrondal, A. (2010). The Cambridge Dictionary of Statistics, Cambridge University Press.

Gonick, L. (1993). The Cartoon Guide to Statistics. Harpe rPerennial.

Moore, David S., "The Basic Practice of Statistics." Third edition. W.H. Freeman and Company. New York. 2003.

Omekara, C. O. and Acha C. K. (2017). Nonparametric Statistics. ISBN: 978-54309-6-7 Prisat Nigeria Limited: Aba.

PRESENTATION SCHEDULE

There are twenty-two units in this course. Each unit represent particular area to be covered in the study. Also, the weekly activities are presented in Table 1 while the required hours of study and the activities are presented in Table 2. This will guide your study time. You may spend more time in completing each module or unit.

COURSE OVERVIEW

This table 1 displays the units, the number of weeks you should take to complete them and the assignment as follows;

The activities in Table I include facilitation hours (synchronous and asynchronous), assignments, mini projects, and laboratory practical. How do you know the hours to spend on each? A guide is presented in Table 2;

Table 2: Required Minimum Hours of Study

S/N	Activity	Hour per Week	Hour per Semester
1	Synchronous Facilitation (Video Conferencing)	2	26
2	Asynchronous Facilitation (Read and respond to posts including facilitator's comment, self-study)	4	52
3	Assignments, mini-project, laboratory practical and portfolios	1	13
	Total	7	91

ASSESSMENT

There are two types of the assessment of the course. The first is the tutor-marked assignments and the second is written examination.

In tackling the assignments, you are expected to apply information, knowledge and techniques gathered during the course. The assignments must be submitted to your tutor for formal assessment in accordance with the deadlines stated in the Presentation Schedule and the Assignments File. The work you submit to your tutor for assessment will account for 30 % of your total course mark.

At the end of the course, you will need to sit for a written examination of three hours' duration. This examination will also account for 70% of your total course mark.

PORTFOLIO

A portfolio has been created for you tagged ―My Portfolio‖. With the use of Microsoft Word, state the knowledge you gained in every Module and in not more than three sentences explain how you were able to apply the knowledge to solve problems or challenges in your context or how you intend to apply the knowledge. Use this Table format:

APPLICATION OF KNOWLEDGE GAINED

Module	Topic	Knowledge Gained	Application of Knowledge Gained

You may be required to present your portfolio to a constituted panel.

MINI PROJECTS WITH PRESENTATION

You are to work on the project according to specification. You may be required to defend your project. You will receive feedback on your project defence or after scoring. This project is different from your thesis.

ASSIGNMENT FILE

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

There are seven assignments in this course which cover:
Assignment 1 - All TMAs' question in Units 1 - 3
Assignment 2- All TMAs' question in Units 4 - 7
Assignment 3 - All TMAs' question in Units 8 - 10
Assignment 4 - All TMAs' question in Unit 11 -14
Assignment 5 - All TMAs' question in Unit 15 -18
Assignment 6 - All TMAs' question in Unit 19-20
Assignment 7- All TMAs' question in Units 21 – 22

Take the assignment and click on the submission button to submit. The assignment will be scored, and you will receive a feedback.

TUTOR-MARKED ASSIGNMENTS (TMAS)

There are seven tutor-marked assignments for this course, all of which you are expected to submit. You are encouraged to work all the questions thoroughly. Each assignment accounts for 8.333% of your total course mark.

Assignment questions for the units in this course are contained in the Assignment File. You will be able to complete your assignments from the information and materials contained in your course material, textbooks and further readings. However, it is desirable in all degree level of education to demonstrate that you have read and researched more widely than the required minimum. You should use other references to have a broad viewpoint of the subject and also have deeper understanding of the subject.

On completion of each of your 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 except for exceptional circumstances.

FINAL EXAMINATION AND GRADING

The final examination will be of three 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-testing, practice exercises and tutor-marked problems you have previously encountered. All areas of the course will be assessed.

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

Finally, the examination will help to test the cognitive domain. The test items will be mostly application, and evaluation test items that will lead to creation of new knowledge/idea.

Table 3 presents the mode you will be assessed.

Table 3: Assessment

S/N	Method of Assessment	Score (%)
1	Portfolios	10
2	Mini Projects with presentation	30
3	Assignments	20
4	Final Examination	40
Total		100

HOW TO GET THE MOST FROM THE COURSE

Distance learning is designed for intellectually mature students who have the capacity for independent study. They are expected to read and work through specially designed study materials at their own pace and at a time and place that suits them. This material serves as a reading material for private study and even goes further to guide students on when to read their books or other materials, and when to undertake tutorials and practical work. Just as a lecturer might give an in-class exercise, the study units provide exercises to be carried out at appropriate points.

Each of the study units adopts a common format. It starts with an introduction to the subject matter and subsequently shows how a particular unit is linked to other units and the course as a whole. Next is a set of learning objectives. These objectives highlight what a student should be capable of doing after working through the study unit.

Students are advised to use these objectives as standards against which to check their appreciation of any unit. Adverse variation between student's knowledge and the specified objectives should be carefully and adequately addressed before moving onto the next study unit. This may make the difference between a lackluster average performance and an excellent result.

The main body of the unit serves as a guide through the required reading from other sources. This will usually be from either the recommended text books or from a Readings section. Some units require the undertaking of computer practical work. The purpose of the computing work is twofold. Firstly, it will enhance the student's understanding of the material in the unit and secondly, it will endow students with practical experience of using programs, which they could well encounter in their work life.

Self-tests are interspersed throughout the units. Working through these tests will help the attainment of the objectives of the unit and prepare students for the assignments and the examination. It is advisable that students practice the self-tests in each study unit. Numerous examples are given in the study units and working through them will be a smart move that acquaints users with the necessary rubrics of the study unit.

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

IMPORTANT INFORMATION;

1. Read this Course Guide thoroughly;
2. Organize a study schedule;
3. Always refer to the `Course overview for more details;
4. Be conscious of the time you are expected to spend on each unit and how the assignments relate to the units;
5. Details of your tutorials and the date of the semester is available at the study centre.;
6. You need to gather all this information in one place, such as your dairy or a wall calendar. Whatever method you choose to use, you should write down your dates for working on each unit;
7. 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;

8. Assemble the study materials. The 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 text books on your desk at the same time;
9. 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 text books or other articles. Use the unit to guide your reading;
10. Up-to-date course information will be continuously delivered to you at the study centre;
11. It will be good to get the Assignment File for the next required assignment before the relevant due date (about 4 weeks before due dates). 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;
12. 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;
13. Study and understand a unit before going to the next unit; when you are confident that you have achieved a unit's objectives, you can then proceed to the next. Proceed unit by unit through the course and try to pace your study so that you keep to schedule. For example, if you should start with unit 1, read and understand the following: Introduction, objective(s), main content, conclusion, summary, tutor marked assignment, references, further readings and other resources before moving on to unit 2;
14. 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;
15. 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 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 before the due date (at least two working days are required). They will be marked by your tutor and returned to you as soon as possible.

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

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

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

ONLINE FACILITATION

There will be two forms of facilitation – synchronous and asynchronous. The synchronous will be held through video conferencing according to weekly schedule. During the synchronous facilitation:

- There will be two hours of online real time contact per week making a total of 26 hours for thirteen weeks of study time.
- At the end of each video conferencing, the video will be uploaded for view at your pace.
- You are to read the course material and do other assignments as may be given before video conferencing time.
- The facilitator will concentrate on main themes.
- The facilitator will take you through the course guide in the first lecture at the start date of facilitation

For the asynchronous facilitation, your facilitator will:

- Present the theme for the week.
- Direct and summarise forum discussions.
- Coordinate activities in the platform.
- Score and grade activities when need be.

- Support you to learn. In this regard personal mails may be sent.
- Send you videos and audio lectures, and podcasts if need be.

Read all the comments and notes of your facilitator especially on your assignments, participate in forum discussions. This will give you opportunity to socialise with others in the course and build your skill for teamwork. You can raise any challenge encountered during your study. To gain the maximum benefit from course facilitation, prepare a list of questions before the synchronous session. You will learn a lot from participating actively in the discussions.

Table 4 is the presentation of the online facilitation of STT 205 Course Plan

TABLE 4: PRESENTATION OF STT 205 COURSE PLAN

MODULES	WEEKS	UNITS			
1	1	Unit 1	Unit 2	Unit 3	-
2	2	Unit 1	Unit 2	Unit 3	Unit 4
3	3	Unit 1	Unit 2	Unit 3	-
4	4	Unit 1	Unit 2	Unit 3	Unit 4
5	5	Unit 1	Unit 2	Unit 3	-
6	6	Unit 1	Unit 2	Unit 3	-
7	7	Unit 1	Unit 2		
	8	Revision			

Finally, respond to the questionnaire. This will help NOUN to know your areas of challenges and how to improve on them for the review of the course materials and lectures.

LEARNER SUPPORT

You will receive the following support:

- Technical Support: There will be contact number(s), email address and chatbot on the Learning Management System where you can chat or send message to get assistance and guidance any time during the course.
- 24/7 communication: You can send personal mail to your facilitator and the centre at any time of the day. You will receive answer to you mails within 24 hours. There is also opportunity for personal or group chats at any time of the day with those that are online.
- You will receive guidance and feedback on your assessments, academic progress, and receive help to resolve challenges facing your studies.

COURSE BLUB

This course presents statistical data collection as an operation of statistical data processing aimed at gathering of statistical data and producing the input object data of a statistical survey. In addition, different sources and methods of data collection are included. Besides, it covers the descriptive handling of statistical data and looks at diagrammatic representation - diagrams, charts and graphs. Furthermore, it provides the process of deducing properties of underlying distributions by analysis of data, known as statistical inference. It also has in it measures of central tendency (location) without leaving the measures of dispersion (spread) which is a lack of uniformity in the sizes or quantities of the items of a group or series. Elementary probability (special probability distributions) and Statistical (probability) distributions are discussed. An x-ray of the basic estimates (point and interval) and principle of hypothesis testing details are considered

MAIN COURSE

CONTENTS

MODULE 1 NATURE OF STATISTICS, STATISTICAL INQUIRIES, FORMS AND DESIGN

In this module, you will be introduced to the Nature of Statistics, Statistical Inquiries, Forms and Design. This module is made up of the following units:

Unit 1 Introduction to Statistics
Unit 2 Sources and Methods of Statistical Data
Unit 3 Statistical Inquiries, Forms and Design: Questionnaire

UNIT 1 INTRODUCTION TO STATISTICS

Unit Structure

1.1 Introduction

The word Statistics has several uses and meanings; it can be used to denote numerical data sets, information from experimental units, and so on. It is also defined differently in various fields like Management Science Statistics, Biostatistics, Agriculture Statistics, Industrial

Statistics, Medical Statistics and Educational Statistics etcetera. As a course of study, however, the word Statistics is used to indicate the proper method of collecting, organizing, analysing, interpreting and drawing conclusion from numerical data sets which may vary from time to time, place to place, trial to trial, person to person, and material to material.

1.2 Intended Learning Outcomes (ILOs)

By the end of this unit, you will be able to know the three main objectives of statistics which are:

- Estimation,
- Prediction, and
- Decision making from analysis of properly collected data.

1.3 Introduction to Statistics

1.3.1 Definition of Statistical Terms:

Statistic is a number resulting from the manipulation of raw data according to certain specified procedures. A statistic is a random variable, which is a function of another random variable(s).

Examples:

Sample mean $\bar{X} = \frac{\sum_{i=1}^{n} Xi}{n}$ 1.1

Sample variance = $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ 1.2

Sample is a subset of a population or universal set.

Population is a complete set of individuals, item or object having some common observable characteristics.

Data is array of numbers or measurements which are collected as a result of observations.

Variable is a characteristic or phenomenon which may take a different value.

Parameter is any characteristics of a population which is measurable. This is a function of population values describing a population sample. An example of parameter is a population value e.g. population mean μ.

Variance is a function that measures the spread between numbers in a data set.

Experiment is any operation carried out and followed by a result.

Random Experiment is an experiment whose outcome may not be the same even though the condition of the experiment may be the same.
The experiment can be conducted repeatedly under the same conditions at different times. Example: Student's score in an examination.

Sample Space is the set of all possible outcomes of a random experiment. It can be finite, countable finite or countable infinite.

A sample space is said to be discrete if it contains finite or countably infinite sample point.

1.3.2 Event

An event is a collection of sample space. It can also be defined as an outcome of a random experiment. Example: number of students that sat for an examination, so the event result would be success or failure.

1.4 Classes of Event

1.4.1 Elementary Event: this is an event consisting of one point.
Example: examination outcome – success or failure; tossing a die once – with scores 1- 6; tossing a coin – with outcome head or tail

1.4.2 Composite Event: this is an event consisting of more than one point. Example: throwing a die more than once and scoring different numbers from 1 to 6.

1.5 Branches of Statistics

The two main branches of statistics are descriptive statistics and inferential statistics.

1.5.1 Descriptive Statistics uses the data to provide descriptions of the population, either through numerical calculations, graphs or tables.

1.5.2 Inferential Statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.

Both of these are employed in scientific analysis of data and both are equally important students of statistics.

1.6 **Variable**: a variable is that characteristics of a population or sample which assumes different values from individuals to individuals within the groups.

Examples of variables are:

- Age of students of final year,
- Weight of animals,
- Height of staff in statistics department.

1.7 Types of variables:

- discrete variables
- continuous variable

1.7.2 **Discrete Variable**: discrete variable is one whose numerical value can have only distinct numbers at a time. Its numbers are usually integer, no measurement.

Examples of discrete variables are:

- Number of rooms in a house,
- Number of people in a city,
- Number of fishes in a pond,
- Fingers of a bananas in a bunch,
- Number of footballers in a field.

Continuous Variable: continuous variable is one whose numerical evaluation at any time can assume a range of values; it depends on accuracy of a measurement.

Example of continuous variables are:

Height said to be 6.15inch could be 5.1866inch with sensitivity balance or 5.778inch on another scale. So between 5 and 6 there is uncountable numbers of values which an element can assume. This type of variable in question determines how some treatments are chosen from experiment and how their effects are graphed.

1.8 The Importance of Statistics

Some basic benefits derivable from statistics are that it:

- Leads to better decision-making across all sectors within an economy; The better the data, the more accurate and trustworthy the decisions that will emanate there from;
- Helps in collecting appropriate quantitative, qualitative and relevant data;

- Decreases risk and results in consistent improvements in results;
- Helps in providing a better understanding and exact description of a phenomenon of nature;
- Enhances quality of formulated policies;
- Teaches people to use a limited sample to make intelligent and accurate conclusions about a greater population;
- Helps in the proper and efficient planning of a statistical inquiry in any field of study;
- Provides tools like tables, graphs, and charts which play a vital role in presenting data used to draw conclusions;
- Many outputs and results are based on statistical models build on statistical concepts like forecasting and prediction;
- To collect the relevant data. Otherwise, it results in a loss of money, time and data using statistical skills;
- It is the pivot of successful trading and investment by businessmen;
- Leads to good recommendations on the way forward.

Self-Assessment Exercise(s)

1.	What is the difference between the following, include three examples of each;
a.	Statistics and Statistic
b.	Elementary Event and Composite Event
c.	Discrete Variables and Continuous Variable
2.	Mathematically illustrate the difference between the sample mean and sample variance.
3.	Define data, sample, population, data, variable, parameter, experiment, random experiment, sample space, event.
4.	List four examples of each type event.
5.	Mention and define two main branches of statistics.

1.9 Summary

Unit 1 Introduces basic terms and definitions and discusses how and when statistics are used in research and in real-life.

The importance of statistics cannot be overemphasized because it is very crucial in all parts of our life. Statistics also plays crucial roles in world development and is applicable everywhere and in everything we do.

1.10 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

Bhattacharyya, G. K., and R. A. Johnson, (1997). Statistical Concepts and Methods, John Wiley and Sons, New York.

Moore, David S., "The Basic Practice of Statistics." Third edition. W.H. Freeman and Company. New York. 2003.

UNIT 2 SOURCES AND METHODS OF STATISTICAL DATA

Unit Structure

2.1 Introduction

Statistics is the study of data collection and analysis. It is mostly used to keep records, calculate probabilities, and provide knowledge. Basically, it helps us understand the world through numbers and other quantitative information.

2.2 Intended Learning Outcomes (ILOs)

By the end of unit 2, you will be able to

- to effectively conduct research;
- to read and evaluate data sets
- to further develop critical thinking and analytic skills.

2.3 Sources and Methods of Statistical Data

2.3.1 Definition of Data Set

Statistical data are characteristics or information, usually numerical, that are collected through observation. Statistical data processing on the other hand is aimed at gathering statistical data and producing the input object data of a statistical survey. It can also be defined as a data set, which is a piece of information collected from population or sample. The singular form of data is datum.

2.3.1.1 Definition of Terms in Data Sets

Array: An array is an ordered arrangement.

Raw Data: raw data is data that has not been processed for use.
It is important to note that information is end product of data processing.

2.3.2 Types of Statistical Data

i. numerical data
ii. non-numerical data
iii. Nominal data

2.3.2.1 Numerical and Non-Numerical Data

Numerical Data*:* these are data reduced to numbers which are purely quantitative depending on the type of variable.

Non-Numerical Data: these cannot be expressed in term of numbers.
Examples non-numerical data are:
Gender: male and female;
Size: small, medium, large;
Colours: white, blue, purple and so on.

Nominal data: They are strictly connected with names and cannot be reduced to numbers.

Examples of nominal data are:
Fishes: cat fish, tilapia, etc.

2.3.2.2 Deductive and Inductive Statistics.

Another way to make a piece of information collected from a sample of a population is to communicate the ideas embedded in numbers through *deductive or inductive statistics*

- ***Deductive Statistics***
  If on the other hand the sample data are only analysed, summarized, and presented without making any reference about the population. We call such phase of statistics deductive statistics.

- ***Inductive Statistics***
  If statistical data, obtained on the basis of a sample selected from a population of units of enquiry, are used to make valid conclusion about the entire population we call such phase inductive statistics or statistical inference.

2.3.2.3 Quantitative and Qualitative Data

Statistical data can also be divided into two: *quantitative and qualitative data.*

- ***Quantitative Data:*** Is any set of data that can be expressed numerically.
  Examples of quantitative data are: age, shoe size, weight, height and so on.

- ***Qualitative Data:*** Is any set of data that cannot be expressed numerically but as a quality. Examples of qualitative data are: sex, occupation, colour, marital status and so on.

After data collection, it is important that data is presented in a way that conveys the information it contains in a clear term. When summarizing large masses of raw data, it is often useful to distribute the data into classes or category and to determine the number of individuals belonging to each class called frequency.

2.3.3 Sources of Statistical Data:

We have two sources:
- *Primary Data and*
- *Secondary Data.*

2.3.3.1 Primary Sources: Data that have been collected by the investigator or his agent directly from the experimental unit. The data

may or may not have been published. Such sources of data are called primary sources.

2.3.3.2 Secondary Sources: Data for which the investigator is not the originator/ initiator but which was collected from other published records, gazettes, books, journals, registers etc. are classified as secondary sources.

2.3.4 Methods of Collecting Data.

For primary data, collection can be done by using any of the following methods;
- Direct observation
- Mail questionnaire
- Personal interviews

2.3.4.1 Direct Observation: This is a method of data collection whereby information is obtained from the units of enquiry by observing them in their natural environment. The investigator observes either the entire population or a sample. This method is usually used in scientific experiment and studies of human health. It is usually applied in collection of data in human and non-human populations.

The participant observer obtains the required information directly by sharing in the activities of the community rather than by relying on the report of the informant or even those of the units of enquiry. Direct observation becomes very necessary when the informants are not able to supply the required information or are likely to give inaccurate answers.

2.3.4.2 Mail Questionnaire: A carefully formulated set of questions may be sent to the respondents by post or carried by field workers or enumerators for their responses.

2.3.4.3 Personal Interviews
This is usually the method of collection of data in social surveys. Here a set of question is prepared and administered on the respondent personally by the interviewers.

Other methods of data collection:
Population census
Experiment: It can be seen as a direct observation. Here events are observed and characteristics of interest are recorded. In performing a particular experiment and recording the observation obtained from such experiments that could be collected in the process. This is a major means of data collection in sciences.

***Focus groups**:* A small discussion group with a group moderator present to keep the discussion focused.

Simulation of data.
Existing data: It can also be seen as a Secondary data since the data are originally collected and then archived or any other kind of "data" that was simply left behind at an earlier time for some other purpose.

Tests: This includes standardized tests such as information on reliability, validity, and norms as well as tests constructed by researchers for specific purposes, skills tests, and so on.

***Sample Survey**:* This is a special survey, it can be a pilot survey or full survey using questionnaires, interviews and other methods.

Self-Assessment Exercise(s)

1. Give two similarities and two differences between;
   i. Numerical and Non-Numerical Data.
   ii. Quantitative and Qualitative Data.
   iii. Deductive and Inductive Statistics.
2. Explain statistical data, array, raw data, cooked data.
3. A single form of data is what?
4. List any ten importance of statistics.
5. Define Statistical data

2.4 Summary

Unit 2 teaches that sample data can be analysed, summarized, and presented without making any reference about the population.

The sources and methods of data collection whereby information is obtained from the units of enquiry were discussed in details.

2.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2821-50-2. Daves Publishers, Uyo.

Alreck, P.L. and Settle, R. (2002). Survey Research Handbook, 2rd edition, McGraw Hill, New York.

Barnes, S. (Ed.) (2005). News of the World, Football Annual 2005–2006, Invincible Press, London, ISBN 0-00-720582-1.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-12-196869-6.

Buglear, J. (2002). Stats Means Business: A Guide to Business Statistics, Butterworth-Heinemann, Oxford, ISBN 0-7506-5264-7.

UNIT 3 STATISTICAL INQUIRIES, FORMS AND DESIGN: QUESTIONNAIRE

Unit Structure

3.1 Introduction

Questionnaire was invented by Sir Francis Galton. A questionnaire can be said to mean a written or electronic survey instrument comprising of a series of questions designed to measure a specific item or set of items. It plays a central role in the data collection process. A well-designed questionnaire efficiently collects the required data with a minimum number of errors. It facilitates the coding and capturing of data. It leads to an overall reduction in the cost and time associated with data collection and processing.

To draw up a good questionnaire, the following questions should be addressed;

1. Why is this survey being conducted?
2. What do I need to know?
3. How will the information be used?
4. How accurate and timely does the information have to be?

Before designing the questionnaire, many decisions have to be made which affect the questionnaire and should be part of the draft plan for the survey.

3.2 Intended Learning Outcomes (ILOs)

By the end of unit 3, you will be able to translate the objectives of the data collection process into a well conceptualized and methodologically comprehensive study.

3.3 Main Content

3.3.1 Definition of a Questionnaire

A questionnaire can be defined as a type of survey instrument handed out in paper form or electronically, normally to a specific demographic population to gather information for a specific purpose. A questionnaire can also be defined as a concise pre-planned set of questions designed to yield specific information to meet a particular need for research information about a pertinent topic. Questionnaire is normally used in study involving human beings and the social environment.

3.3.2 Qualities of a Good Questionnaire

The design of a questionnaire will depend on whether the researcher wishes to collect exploratory (qualitative) information for the purpose of better understanding or the generating of hypothesis on the subject or quantitative information (to test specific hypothesis that have previously been generated). The following are some qualities of a good questionnaire:

1. A well-designed questionnaire should meet the research objectives.
2. It should obtain the most complete and accurate information possible.
3. A well-designed questionnaire should make it easy for respondents to give necessary information and for the interviewer to record the answer. It should also be arranged so that sound analysis and interpretation are possible.
4. It would keep the interviewer brief and straight to the point and be arranged to make the respondent(s) give good responses throughout the interview.

3.3.3 Types of Questionnaires

- ***Exploratory Questionnaire:*** If the data to be collected is qualitative or is not to be statistically evaluated, it may be that no formal questionnaire is needed.

- ***Formal Standardized Questionnaire****:* If the researcher wants to test and quantify hypothesis then the data is to be analysed statistically. Here a formal standardized questionnaire is designed and such questionnaire is generally characterized by:

- Precise wording and order of questions, to ensure that each respondent receives the same stimuli
- Prescribed definition or explanation for each question consistently and can answer respondents' requests for clarification if they occur.
- Prescribed response format to ensure rapid completion of the questionnaire during the interviewing process.
- Concise set of questions.

3.3.4 Parties to a Questionnaire

In a questionnaire, three people are involved.
- Enumerator
- Respondent
- Enumeration/Observation Unit

3.3.5 Enumerator: This is a name sometimes given to an investigator. An investigator is the person at head of the whole research effort. It may be government, group or organization.

3.3.6 Respondent: This is the person that answers the questions in the questionnaire.

3.3.7 Enumeration /Observation Unit: This is the act of mentioning a number of things one by one.

3.4 Design of Questionnaire

The use of questionnaire is the most satisfactory method of collecting primary data. There are no hard-and fact rules on how to design a questionnaire but there a number of points that can be kept in mind. The following guidelines are necessary for the design of a good questionnaire.

- Number of Questions should be as few as possible and direct on the topic being investigated, that is, avoid too many questions and repetition where possible.
- Order of Questions should follow a logical sequence and should be well arranged and ordered.
- A well designed questionnaire should meet the objectives.
- Questions should be simple, clear and unambiguous for more meaningful responses particularly for complicated & ambiguous questions.

- Questions which demand definite answers (close ended), should be used rather than those which demand various answers (open ended).
- Leading questions should be avoided.
- Confidential questions which pry into private or personal affairs should be avoided.
- Foot notes where necessary as a guide to respondent should be incorporated.
- Questions that provoke strong feelings & probe into respondent privacy should be avoided and when such questions are unavoidable, explanatory notes should be used to facilitate the respondent response.

3.5 Forms of Questionnaire

- ***Dichotomous:*** This refers to questions that have only two answers. Example: Yes or No, Male and Female, etc.
- ***Multiple Choices:*** This involves more than two options. The respondent is expected to choose only one of these options.

Example:

1. Through which source did you hear about bank consolidation?

A. Radio
B. Television
C. Newspaper
D. Internet
E. Friends.

- ***Open ended:*** Here a questionnaire is designed in such a way that the respondent has to express his or her view on a particular issue at hand**.**

3.6 Components of a Good Questionnaire

- Letter of Introduction: Here you first introduce yourself to the target recipients of your questionnaire.
- Question on the respondent's, religion, background place of residence, age, sex, marital status, educational attainment, etc.
- The complementary close; At this point, you are expected to express praise or admiration.

3.7 Advantages and Disadvantages of Use of Questionnaire

Advantages of Use of Questionnaires

i. The respondent can fill the questionnaires without pressure from the interviewers.
ii. Respondent may likely respond more to confidential questions.

iii. It is less expensive
iv. Respondent can be reminded of any delay through letter or telephone

Disadvantages of Use of Questionnaire

i. Low rate of response
ii. Gross irregularities in filling the questionnaires
iii. Bias may follow a number of responses

Self-Assessment Exercise(s)

1. Define of a questionnaire and write three qualities of a good questionnaire.
2. State the eight steps involved in designing a questionnaire.
3. Distinguish between dichotomous, multiple choice and Open ended questionnaire.
4. List the components of a good questionnaire.
5. Write three advantages and three disadvantages of questionnaire.

3.8 Summary

Unit 3, discusses in the details of how to translate the objectives of the data collection process into a well conceptualized and methodologically comprehensive study using questionnaires.

In this unit, you have learnt to translate the objectives of the data collection process into a well conceptualized and methodologically comprehensive study.

3.9 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Berenson M.L. and Levine D.M. (1996) Basic Business Statistics, Prentice-Hall, Englewood Cliffs, New Jersey.

Buglear, J. (2002). Stats Means Business: A Guide to Business Statistics, Butterworth-Heinemann, Oxford, ISBN 0-7506-5364-7.

MODULE 2 PRESENTATION OF STATISTICAL DATA

Module 2 covers the presentation of statistical data: descriptive handling of statistical data frequency, diagrammatic representation, ratios, percentages and numbers from unit 1 to unit 4.

Unit 1 Descriptive Handling of Statistical Data
Unit 2 Frequency
Unit 3 Diagrammatic Representation
Unit 4 Ratios, Percentages and Random Numbers

UNIT 1 DESCRIPTIVE HANDLING OF STATISTICAL DATA

Unit Structure

1.1 Introduction

We shall in general, discuss how to make tables, charts and diagrams from data which are observed over time and space, while grouping of data from experiment and surveys into frequencies will be discussed later. In both cases, the way it is done will depend on the scale (level of measurement). Most elementary things we are expected to know about collection of numerical data will be expatiated on.

1.2 Intended Learning Outcomes (ILOs)

The major objective of this unit is reducing or simplifying the details given in a mass of data into such form that the main features may be brought out to make the assembled data easily understood.

1.3 Main Content

1.3.1 Statistical Tabulation

Statistical tabulation is a systematic and logical presentation of numeric data in rows and columns to facilitate comparison and statistical analysis. In other words, the method of placing organised data into a tabular form is called tabulation. In tabulation, classification, forms the basis of reducing and simplifying the details given in a mass of data into such form that the main features may be brought out to make the assembled data easily understood. This is done on a statistical table where the data may be arranged in columns and/or rows. The purpose of tabulation is to condense and thereby facilitate comparison of data. Tabulation is the first kind of element any summary work in statistical investigations can take. Most published data usually come in this form and is one of the most popular methods of making data more comprehensible.

1.3.2 Kinds of Tabulations

Broadly, there are two kinds of tabulations

- Tables, presented as the values are recorded in particular overtime and space. If overtime, they are called time series and if over space, they are called cross sectional data.
- Tables, presented such that the values have been categorized with their frequencies and each category recorded. The resulting tables are called frequency table.

1.3.3 Types of Tables

Basically, there are three types of tables use in presenting statistical work. They are:

- Source or reference table
- Working table
- Summary or text table

The source table is one on which further analysis is based.

The working table as the name suggests is a sheet on which initial calculations are drawn before the final tables are arrived at summary or text tables. This is usually found in books either to support the argument presented or to have easy reference.

A summary table is usually a derived table that is modified to suit a particular purpose.

1.3.4 Characteristics of Tables

A statistical table has some general features no matter the purpose for which the table is constructed. Some of these are:

- A general title which is a brief explanation of what the table is about;
- Column title or caption to show order of classification along the columns;
- A row title or stub to show order of classification along the row;
- A source note at the bottom which gives the source of information contain in the table and possibly a short note on the data;
- An indication of the units in which the data in the table is given.

All these characteristics are presented in table I is an example of how a table is to be presented. The basic rule is to avoid a complicated table or give too much information in any one table.

1.3.5 Preparation of a Table

- Sketch the skeleton of the table in blank.
- Arrange of data according to some basis classification or some customary order, that is, in order of magnitude or alphabetical order.
- Units should be carefully and completely stated.
- The layout of the table is essential. The spacing, type and ruling can be varied with advantage.

Table 3.1 below is provided as an example to illustrate these characteristics.

Table 3.1: The GDP of Export and Import (2007-2014)

YEAR	GDP(₦)	Import(₦)	Export(₦)
2007	47619.66	12839.60	11023.3
2008	49069.28	10770.40	8206.40

2009	43107.38	8903.70	7402.40
2010	49621.43	7178.30	9088.00
2011	67908.44	7062.60	11720.80
2012	69146.99	4983.60	8920.60
2011	104212.84	17861.70	30360.60
2014	119084.30	21444.70	31192.80

Source: Central Bank of Nigeria Statistical Bulletin, 2014.

1.3.6 Advantages and Disadvantages of Use of Tables

Advantages of Use of Tables

- The result is likely to be more comprehensible, when a mass at data is tabulated.
- Formerly obscure relationships become easier to see.
- A table explains more, in a very little time, than pages of an explanatory essay.

Disadvantages of Use of Tables

- Individual values become unimportant, when data are tabulated.

SELF-ASSESSMENT EXERCISE(S)

1. What are the two kinds of tabulations?
2. Explain the three types of tables.
3. What are the five Characteristics of Tables?
4. Write steps used to preparation of a Table.
5. List three advantages and one disadvantages of a Table.

1.4 Summary

When data are collected and put into numerical form, they do not seem to be meaningful until they are summarized, presented in tables, grouped into categories or frequencies, or prepared as charts and diagrams and summary calculations made which is what we achieved in unit 5.

They main focus of this unit is to simplifying the details given in a mass of data into such form that the main features may be brought out to make the assembled data easily understood.

1.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Berenson M.L. and Levine D.M. (1996) Basic Business Statistics, Prentice-Hall, Englewood Cliffs, New Jersey.

Buglear, J. (2002). Stats Means Business: A Guide to Business Statistics, Butterworth-Heinemann, Oxford, ISBN 0-7506-5364-7.

UNIT 2 FREQUENCY

Unit Structure

2.1 Introduction

In a set of data, the number of times a particular value occurs is called frequency of that value.

2.2 Intended Learning Outcomes (ILOs)

The major outcome is to covers one of the basic uses of statistics, which is organizing raw data into simpler, more useful and understandable form by creating a frequency distribution. This chapter also introduces how to graph statistical information.

2.3 Main Content

2.3.1 Frequency Distribution

For large mass of data, the frequency distribution of frequency table is a tabular arrangement of data by classes together with its corresponding class frequencies.

Example: The following are the ages of children in Christ the King Choir: 12, 8, 9, 12, 15, 10, 9, 12, 10, and 12.

Table 3.1: Ages of Children in Christ the King Choir

Age	Tally	Frequency
8	I	1
15	I	1
9	II	2
10	II	2
12	IIII	4

Table 3.1 is generally called a frequency distribution or frequency table. Therefore, frequency distribution can be defined as a table of collection of data in a particular order with frequency attached to its value or group of values.

2.3.2 Group Frequency Distribution

Sometimes, data are too much and becomes practically impossible to put them in a normal table as in the example above. They are then put into specified groups known as class interval. The frequency can conveniently be known through group arrangement, that is, some set of numbers belonging to a particular group of class. Class size is the Number of values in a class interval.

Table 3.2: Frequency Distribution between 21 to 100

Class interval	Tally	Frequency
21-30	~~IIII~~ IIII	9
31-40	~~IIII~~ III	8
41-50	~~IIIIII~~ IIII	14
51-60	~~IIIIIIIII~~ III	18
51-70	~~IIIIIIIIIIII~~ I	21
71-80	~~IIIIIIIII~~	15
81-90	~~IIIIII~~	10
91-100	~~IIII~~	5
Total		100

2.3.2.1 Class Interval and Class Limits

Any symbol defining a class such as 21-30, 31- 40 and so on, as in table 2.2 is called the class interval. In the first class interval, the value 21 is the lower class limit while 30 is the upper class limit. A class interval which has either lower or upper class limit is said to be an open class interval.

2.3.2.2 Class Boundary/True Class Limit

They are chosen such that no observation of value in the raw data being grouped takes its value. Class boundaries are obtained in practice by taking the average of upper limit of a class and the lower class limit of the next class, and also the average of the upper limit of the first class and lower class limit of the second class. Once a class boundary value is obtained, class boundaries of other class can easily be obtained by adding or subtracting ± 0.5 from the class size as shown in table 3.3.

Table 3.3: Class Boundaries of a Data set

Class interval	Class Boundaries	Frequency
44-48	43.5-48.5	8
49-53	48.5-53.5	19
54-58	53.5-58.5	10
59-63	58.5-63.5	7
64-68	63.5-68.5	5

2.3.2.3 Class Size or Class Width of a Class Interval

A class size or class width of a given class interval, C, is the number of element contained in the class. It is obtained by taking the difference between the lower (upper) limit of a class and the lower (upper) limit of the next higher class.

Example:$C = 53.5 - 48.5 = 5 \; or \; C = 63 - 58 = 5$

2.3.2.4 The Class Mark

The class mark X or class midpoint is obtained as the average of the lower and upper class limit (or boundaries). For unequal class intervals or open classes we still obtain the class mark.

Example: for the Class mark;

$$class\ mark = \frac{44 + 48}{2} = 46$$

2.3.3 Relative Frequency

The relative frequency is obtained as a ratio of the class frequency of all the classes. It can also be expressed as a percentage. A table showing the class and relative frequency is known as the relative frequency distribution.

2.3.3.1 Cumulative Frequency

The cumulative frequency of a given class is obtained by the addition of all class frequencies up to that particular class starting from the lowest class.

2.3.3.2 Cumulative Frequency Distribution

This is the tabular arrangement of data which shows the number of observations less than a given upper class boundary or less than or equal to a given upper class limit.

Using the data in table 3.3, we can generate the cumulative frequency table.

Table 3.4: Table showing less than cumulative frequency distribution of Students Ages.

Less than 12.5	8
Less than 17.5	27
Less than 21.5	37
Less than 27.5	44
Less than 32.5	50

2.4 Histogram

Histogram is a graphical representation of frequency distribution. It has no gap between the bars. To plot the histogram, you plot the frequencies against the class boundaries.

It is constructed by erecting bars on class boundaries of any given class interval of a grouped frequency distribution. The height of each bar is proportional to the frequency of that class and the centre of each bar is the class mark interval. The width of each bar corresponds to the size of the class interval.

In a histogram the vertical axis is the frequency axis, while the horizontal axis n is the class boundary. The reason for using class boundary is to ensure continuity of the graph and eliminate gaps between bars.

2.5 Frequency Polygon

This is a line graph of frequency plotted against class mark. It can be obtained by connecting or joining t he midpoint of the tops of the bars in the histogram. It is customary that both ends of the frequency polygon

be closed to the horizontal axis by considering the next lower and higher class mark by having zero frequencies.

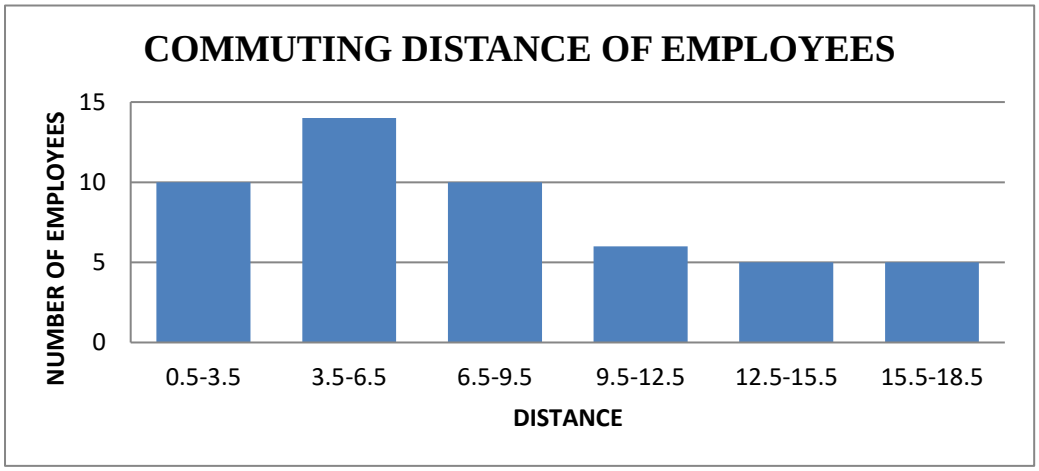
2.6 Cumulative Frequency Polygon (Ogive Curve)

This polygon is obtained by plotting the cumulative frequency less than any upper class boundary against the upper class boundary. It is a very important graph use to locate the quantities, deciles and percentiles.
Example: Suppose a researcher is interested in the number of kilometres that the employees of a large department store travelled to work each day. The researcher may first have to collect the data by taking a sample of 50 employees and the approximate distance of the store from their homes. After doing that the researcher organizes the data in a frequency distribution as follows in table 3.5:

Table 3.5: The number of kilometres that the employees travelled to work each day

Class limits (in miles)	No of employees
1 – 3	10
4 – 5	14
7 – 9	10
10 – 12	5
11 – 15	5
15 – 18	5

For this frequency distribution to be useful the class limits will have to be changed to class boundaries and we make columns for the midpoints,



lower boundaries and cumulative frequencies will have to be added.

Figure 3.1: frequency distribution plot for the number of kilometres that the employees travelled to work each day.

Table 3.5: The cumulative frequency of the employees travelled to work each day

CLASS	MID POINT	LOW BOUND	FREQUENCY	CUM. FREQ.
0.5-3.5	2	0.5	10	0
3.5-5.5	5	3.5	14	10
5.5-9.5	8	5.5	10	24
9.5-12.5	11	9.5	5	34
12.5-15.5	14	12.5	5	40
15.5-18.5	17	15.5	5	45

FREQUENCY POLYGON

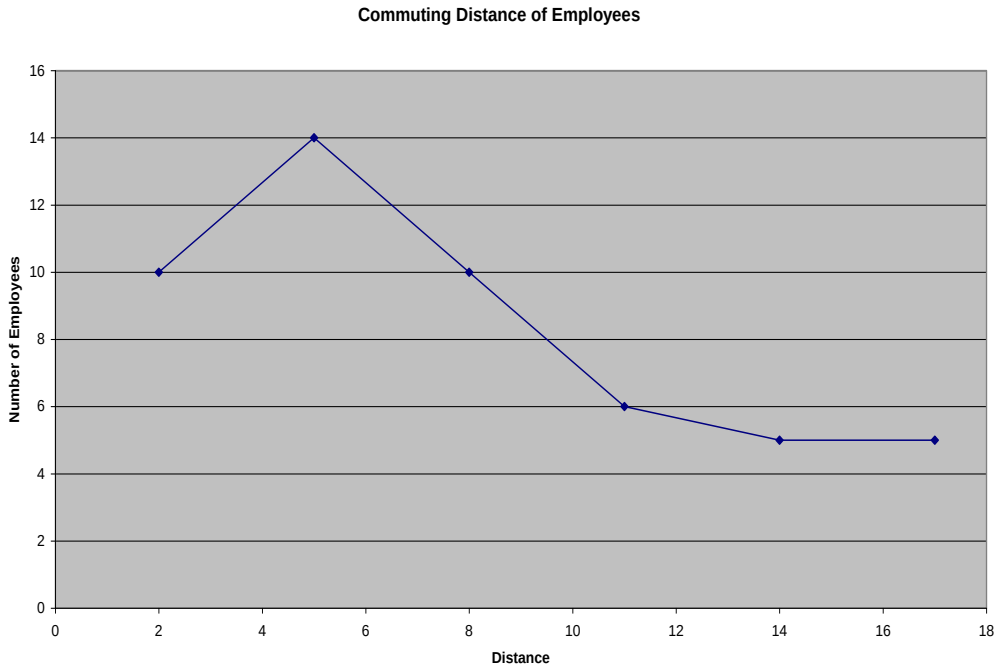


Figure 3.2: The frequency polygon of the employees of a large department store travelled to work each day

Cumulative Frequency Polygon/Curve (Ogive)

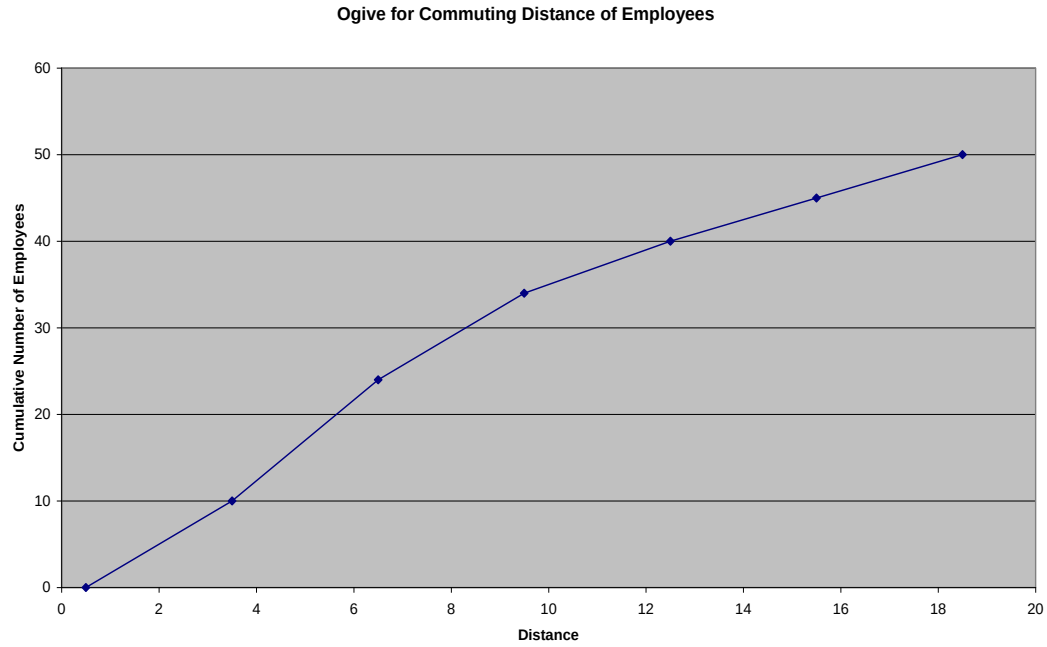


Figure 3.3: The cumulative frequency of the employees of a large department store travelled to work each day

SELF-ASSESSMENT EXERCISE(S)

1. What do you understand by each of the following: Frequency,
2. Frequency Distribution, Group Frequency Distribution?
3. Define Class Size, True Class Limit and Class Width of a Class Interval.
4. Differentiate between Class Interval and Class Limits; Class Boundary and Class Mark.
5. Distinguish between Relative Frequency, Cumulative Frequency and Cumulative Frequency Distribution.
6. Suppose a researcher is interested in the number of kilometres that the employees of a large department store travelled to work each day. The researcher would first have to collect the data by asking a sample of about 50 employees the approximate distance the store is from their homes. After doing that, the researcher organizes the data in a frequency distribution as follows in table 3.7. Obtain the corresponding Histogram, Frequency Polygon and Cumulative Frequency Polygon/Curve.

Table 3.7: The number of kilometres that the employees of a large department store travelled to work each day

Class limits (in km)	No of employees
1 – 3	5
4 – 6	14
7 – 9	5
10 – 12	6
11 – 15	10
16 – 18	10

2.7 Summary

This unit covers one of the basic uses of statistics, which is organizing raw data into simpler and more useful and understandable by creating a frequency distribution, Histogram, Frequency Polygon and Cumulative Frequency Polygon/Curve**.**

This chapter covers one of the basic uses of statistics, which is organizing raw data into simpler, more useful and understandable form by creating a frequency distribution and also introduces how to graph statistical information

2.8 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Buglear, J. (2002). Stats Means Business: A Guide to Business Statistics, Butterworth-Heinemann, Oxford, ISBN 0-7506-5364-7.

Dodge, Y. (2008). The Concise Encyclopaedia of Statistics. Springer. Everitt, B. S.; Skrondal, A. (2010). The Cambridge Dictionary of Statistics, Cambridge University Press.

Gonick, L. (1993).The Cartoon Guide to Statistics. Harper Perennial.

UNIT 3 DIAGRAMMATIC REPRESENTATION DIAGRAMS, CHARTS AND GRAPHS

Unit Structure

3.1 Introduction

Here, the statistical data or information will be depicted in diagrams, pictures, charts, graphs of different types. This will make information simple, clear and unambiguous. It will also help convey the message from the information faster**.**

3.2 Intended Learning Outcomes (ILOs)

Unit 6 introduces how to graph statistical information to achieve a desired output.

3.3 Main Content

3.3.1 Pictorial Diagram

The simplest and most obvious way of presenting information is by means of pictorial figures or designs which are directly related to the item with which the statistical data are collected. There are many forms of pictorial diagrams.

3.3.2 Pictogram

Pictogram is simply a representation in form of pictures as will be seen in example below.

Time	Population (in millions)
2015/16	30
2016/16	60

If one unit of picture represents 10 million people, 30 million will have 3units pictures and 60 million will have 6 units as shown in figure 6.1

2019/20

2020/21

Figure 3.2: Pictogram: Population of Nig. 2019/20 and 2020/21

Some of the essence of putting data into some pictorial or diagrammatic form are:

- It aids the understanding of the tabulated data.
- Diagrams like statistical tables are designed not only to catch the eye but also to convey information. (All diagrams must be properly scaled so that they do not lose the important objective of information).
- They convey broad relationships to the eye even though the intricate details are likely to be blurred.
- Pictorial diagrams are popularly used by journalist in newspapers and by advertisers to show relationships between variables.

3.3.1.2 Block Diagram

A block broken down into the various sections that make up the data. The steps involve are as follows;

Step I: Choose a convenient scale that can show clearly the total figure.

Step II: Use the scale chosen in step I to find the size of each section and the total.

Step III: Draw a block of the size of the total figure and mark each section of it.

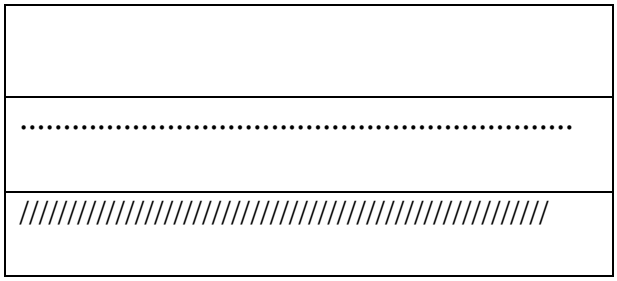


Figure 3.3 represent block diagram (Scale: 1cm represents N100 million)

3.3.2.1 Scatter Diagram

A scatter diagram is a graph in which each plotted point represents an observed pair of values for the independent and dependent variables. The value of the independent variable X is plotted in respect to the horizontal axis and the value of the dependent variable Y is plotted in respect to the vertical axis.

If X and Y denote the two variables under consideration, a scatter diagram shows the location of points (X, Y) on a rectangular coordinate system. If all points in this scatter diagram seem to lie near a line, the relationship is linear.

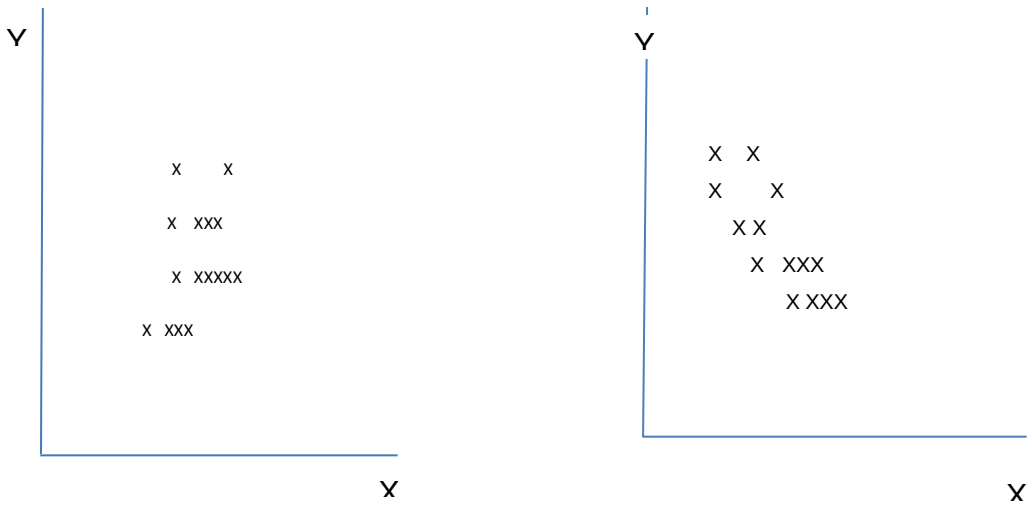


Figure 3.4: Illustration of a scatter diagram

3.3.3 Graph

A graph shows the relationship between two or more variables; usually the variable of interest is called dependent variable and the factor on which the variables of interest depends is called independent variable.

3.3.3.1 Characteristics of a Graph

Students are likely to be familiar with the principle underlying the construction of a graph in mathematics. They should be familiar with the importance of y-axis which is at the vertical axis along which the variable of interest is represented and the x-axis (horizontal axis) along which the independent variable or factor is represented. They are also familiar with the choice of scale which is used to demonstrate the important features of the graph. The scale should always start from zero but where this is not possible there must be some indication to show this. This is very important to draw all graphs. The position of any point on a graph is located by the co-ordinate of the point.

A good graph must satisfy the following points

- The layout. The space allocated for the graph must be large enough and must be fully used. The choice of the scale is important if more than one curve is presented on a graph, the curve must be distinguished. The title must be clear and concise.
- State clearly as in the case of a table the following;

- the title
- the scale
- the unit of increase
- the source of the data

The form of diagram to be used depends on the form of data to be presented.

3.3.3.2 Line Graph:

The graph below depicts simple illustration of a line graph

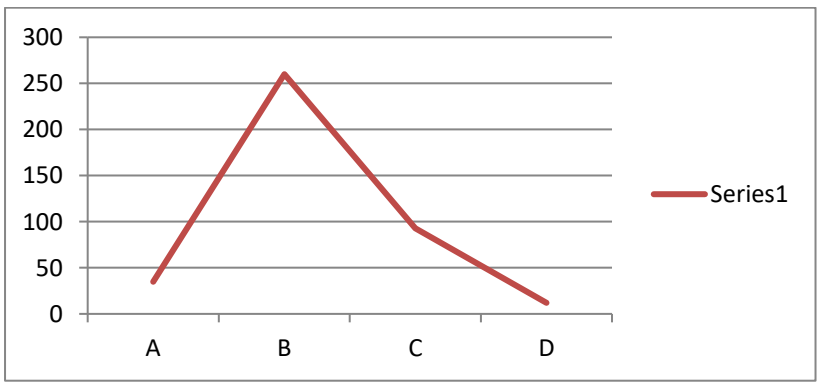


Figure 3.5: illustration of a line graph

3.3.3.3 Series Graph

The graph below shows when there are many line graphs in one plot and this type of graph is called series graph.

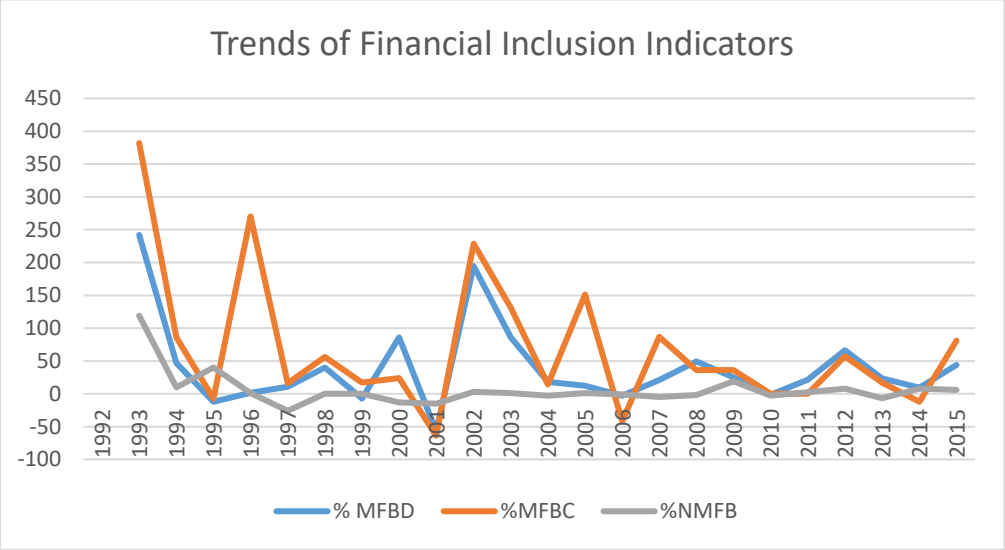


Figure 3.6: illustration of a series graph

3.3.4 Charts

3.3.4.1 Pie Chart: A pie chart is simply a circle divided into sectors. It is a circular method, where each member of the data is represented in a circle. In fact, this circle represents the total of the data being represented and each sector is drawn proportionally to its relative size. The angles of the sectors are proportional to the frequencies of the numbers, objects or classes.

Discussion

Steps for Constructing a Pie Chart

Step 1: Calculate the percentage of each type of item of the total.

Step II: The total of all items is represented by 360^0 which is the angle in a circle.

Step III: Draw a circle of reasonable size and mark the angle obtains in step II. Mark each section/sector differently to show the various compositions.

Example on pie chart:

In a survey concerning public education, 400 school administrators were asked to rate the quality of education in the Nigeria. Their responses are summarized in Table 3.2. Construct a pie chart for this set of data.

Solution:
To construct a pie chart, assign one sector of a circle to each category. The angle of each sector should be proportional to the proportion of measurements (or relative frequency) in that category. Since a circle contains 360°.

Table 3.2: Summary on survey the quality of education in the Nigeria

RATING	FREQUENCY	RELATIVE FREQ.	PERCENT	ANGLE
A	35	35/400 = .09	9%	.09X360^0=32.4^0
B	260	260/400= .65	65%	234^0
C	93	93/400=.23	23%	82.8^0
D	12	12/400 =.03	3%	10.8^0
TOTAL	400	100%		360^0

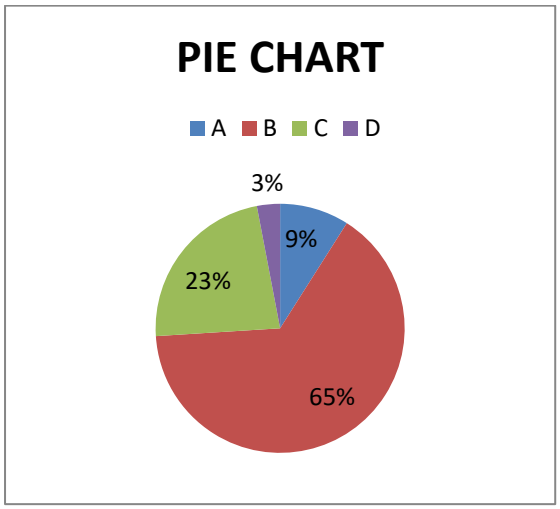


Figure 3.6: Survey on the quality of education in the Nigeria

3.3.4.2 Band Curve Chart

This is a very useful chart which is likely to show much of the information of a set of data. The chart consists of the line graphs of the various items of the total but drawn one above the other. The area between the lines is shaded so that the finished chart will have the appearance of a series of bands. The chart is very suitable to represent data in cumulative flows.

The following steps are taken

Step I: Cumulate the data
Step II: Plot each column on the same graph sheet and connect the point as for a line graph.
Step III: Shade each band differently from another band.
The lines forming the bands are also graphed.

 **Discussion**

3.3.4.3 Bar Chart

Bar Chart is a block diagram in which the length of the bars is proportional to the quantities they represent. The bars must be spaced equally and the width of the bars must be the same. It can be drawn vertically or horizontally.

3.3.5 Types of Bar Chart

3.3.5.1 Simple Bar Diagram/Chart

Simple Bar Chart: simple bar chart is a block diagram in which simple bars are used to denote magnitudes. This is another name for a block diagram and there is no difference between the two, in the sense that the same steps are involved in their construction. They are used for comparison over many years. The diagram is drawn according to scale that represent each year with the bar separated by gaps. The time is shown on the horizontal axis and the other variable (quantity or value) is shown on the vertical axis.

Example on simple bar diagram/ chart:

Table 3.1: Exports of Pakistan (in US $ million)

Year	Exports
1948	118
1951	406
1961	368
1961	683
1981	2958
1991	6168
2001	9202
2005	14410

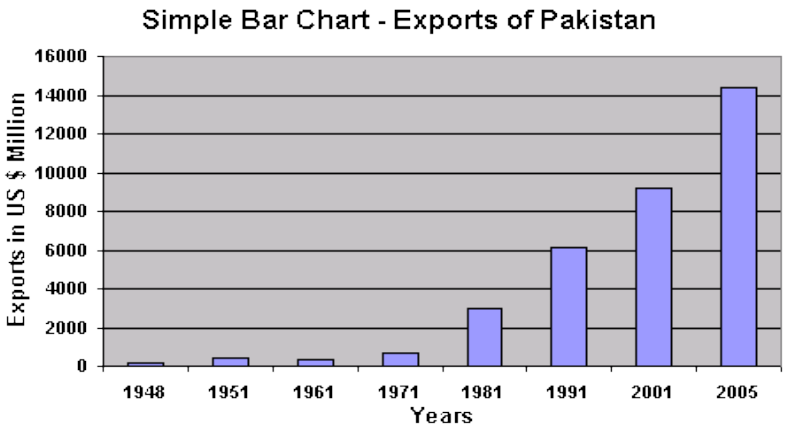


Figure 3.7: Simple Bar Chart- Exports Of Pakistan

Example on Bar Chart for Group data using Table 3.6

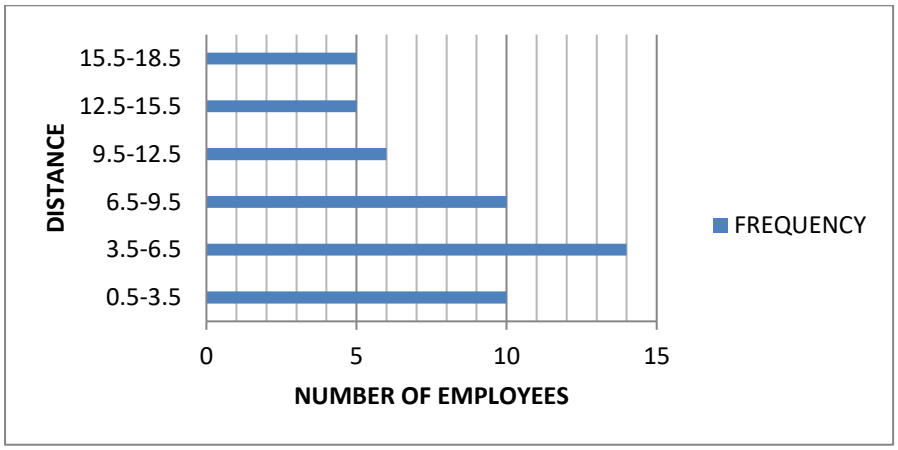


Figure 3.8: Bar Chart for the number of miles that the employees of a large department store travelled to work each day

3.3.6 Component Bar Chart

Component bar chart is a chart that shows part of a whole. It is a chart that is used to represent data in which the total magnitude is divided into components. The simple bar chart is also called simple bar diagram. In a simple bar chart the line, bars, or columns are of equal thickness but the length is varied in proportions to the difference in figure presented. The bars can be vertical or horizontal but where periods of time are concerned like fig. 5. It is usually to show such periods along the horizontal axis. A more complicated type is the component bar chart and multiple bar chart. They are divided into sections. Each section corresponding in size to the magnitude of the item it represent.

 **Discussion**

Steps for constructing Component Bar Chart

Step I: From the raw data, obtain the highest and the lowest values and then the range = The highest value – The lowest value.

Step II: Determine the no. of the data range at class width C, for equal class interval. Number of Class = Range

Step III: Obtain the first class interval or class boundary which contains the lowest value. Other classes are obtain by adding the class size, C, to the lower and upper limit of the first class until we get a class which will absorb the highest value.

Step IV: We then tally the data into the classes they belong to obtain the corresponding class frequencies.
In addition to the steps, the following must be considered;

- This chart consists of bars which are sub-divided into two or more parts.
- The length of the bars is proportional to the totals.
- The component bars are shaded or coloured differently.

Example:
(a) <u>Component Bar Chart:</u>

Table 3.3: Current and Development Expenditure – Pakistan (All figures in Rs. Billion)

Years	Current Expenditure	Development Expenditure	Total Expenditure
1988-89	153	48	201
1989-90	166	56	212
1990-91	196	65	261
1991-92	230	91	321
1992-93	262	66	348
1993-94	294	61	365
1994-95	346	82	428

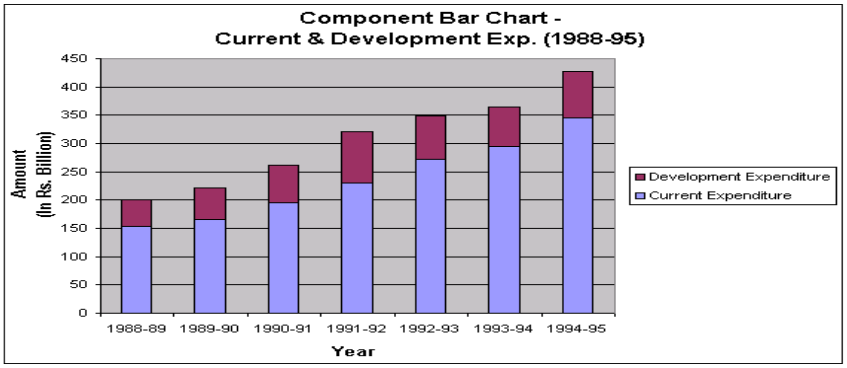


Figure 3.9: Current and Development Expenditure – Pakistan

3.3.7 Percentage Component Bar Chart:

- Component bar charts may also be drawn on percentage basis by expressing the components as percentages of their respective totals.
- All the bars are of equal length showing the 100%. These bars are sub-divided into component bars in proportion to the percentages of their components.

Table 3.4: Areas under Crop Production (1985-90)

Year	Wheat	Rice	Others	Total
1985-86	6403	1863	1926	11192
1986-86	6606	2066	1906	11668
1986-88	6308	1963	1612	10883
1988-89	6630	2042	1966	11638
1989-90	6659	2106	1960	11836

Table 3.5: Percentage Areas Under Production ('000 hectares)

Year	Wheat	Rice	Others	Total
1985-86	66.2%	16.6%	16.2%	100%
1986-86	66.0	16.6	16.3	100
1986-88	66.2	18.0	14.8	100
1988-89	65.9	16.4	16.6	100
1989-90	65.6	16.8	16.6	100

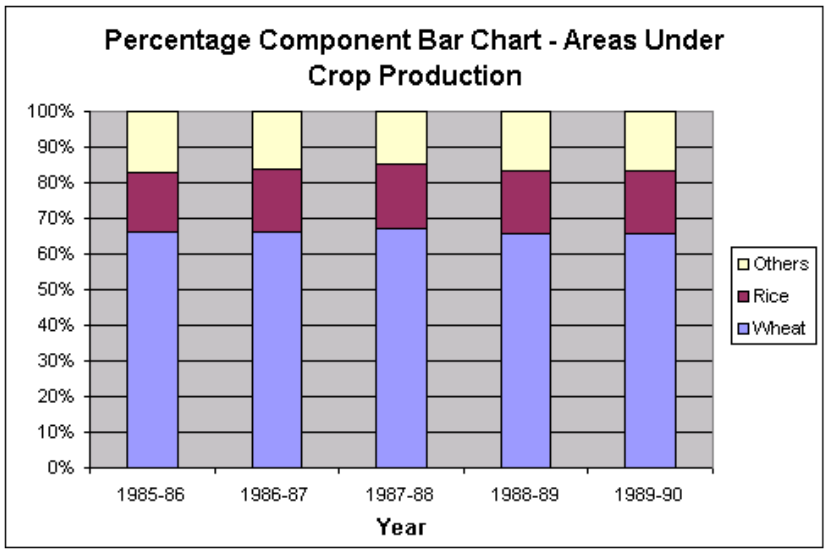


Figure 3.10: Percentage Areas under Production

3.3.8 Multiple Bar Chart

- Multiple bar chart is an extension of simple bar chart.
- Grouped bars are used to represent related sets of data.

For example, imports and exports of Nigeria are shown in multiple bar chart.

Each bar in a group is shaded or coloured differently for the sake of distinction.

Table 3.6: Imports and exports of a Nigeria (1982 -88)

Years	Imports	Exports
1982-83	68.15	34.44
1983-84	66.61	36.33
1984-85	89.68	36.98
1985-86	90.95	49.59
1986-86	92.43	63.35
1986-88	111.38	68.44

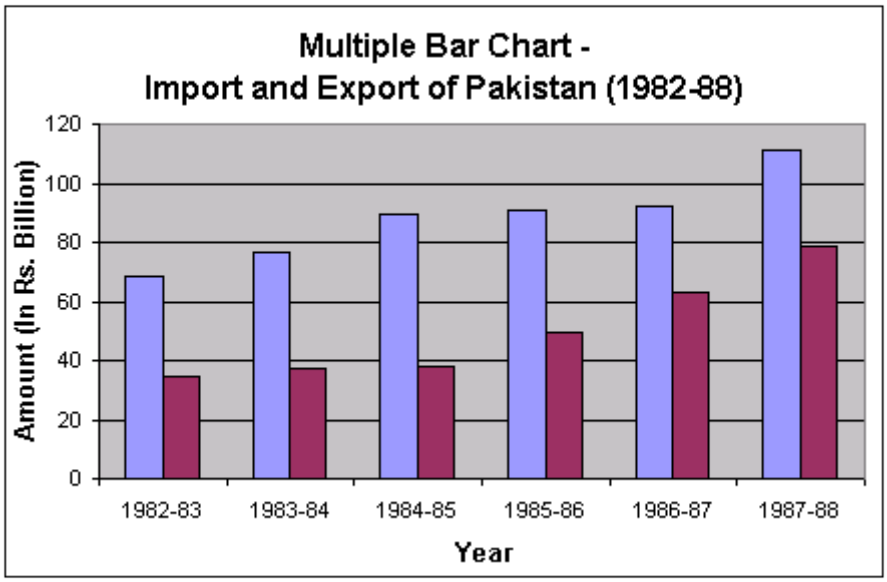


Figure 3.11: Imports and exports of Nigeria (1982 -88)

Advantages of Diagrammatic Representation are:

i. It aids the understanding of the tabulated data.
ii. Diagrams like statistical tables are designed not only to catch the eye but also to convey information. All diagrams must be properly scaled so that they do not lose the important objective of information.
iii. They convey broad relationships to the eye even though the intricate details are likely to be blurred.
iv. Pictorial diagrams are popularly used by journalist in newspapers and by advertisers to show relationships between variables.

Self-Assessment Exercise(s)

1. What is a Block Diagram and what are the steps for constructing it?
2. Define the following: Line Graph, series graph, Pie Chart, Band Curve Chart, Component Bar Chart, Percentage Component Bar Chart, Multiple Bar Chart and Scattered Diagram.
3. What do you understand by pictorial diagram, Pictogram, Simple Bar Diagram/Chart, Bar Chart for ungroup data, Bar Chart for Group data, graph, and charts?
4. What are the Characteristics of a Graph?
5. What the Steps in Constructing a Pie Chart, Band Curve Chart and Bar Chart?

3.4 Summary

The simplest and most obvious way of presenting information is by means of pictorial figures or designs which are directly related to the item from which the statistical data are collected.

This unit covers the most obvious way of presenting information by diagrammatic representation diagrams, charts and graphs.

3.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Buglear, J. (2002). Stats Means Business: A Guide to Business Statistics, Butterworth-Heinemann, Oxford, ISBN 0-7506-5364-7.

Dodge, Y. (2008). The Concise Encyclopedia of Statistics. Springer. Everitt, B. S.; Skrondal, A. (2010). The Cambridge Dictionary of Statistics, Cambridge University Press.

Gonick, L. (1993). The Cartoon Guide to Statistics. HarperPerennial.

UNIT 4 RATIOS, PERCENTAGES AND RANDOM NUMBERS

Unit Structure

4.1 Introduction

Unit 4 compares quantities effectively and checks what the values that we get can tell us about the magnitude of these quantities.

4.2 Learning Outcomes (ILOs)

To give comparative facts on the component parts in relation to the whole and also relates observations in the same variables at different time intervals at different places etc.

4.3 Main Content

4.3.1 Ratios: Ratios are fractions which express variation in the data irrespective of actual or absolute size of the data while percentages are ratios express with hundred (100) as the denominator; usually not written.

Example3.1: If the total expenditure is N36 out of which N18 was spent on food, then the ratio of food to total expenditure is

- Ratio. $\frac{18}{36} = \frac{1}{2}$, This kind of ratio is usually referred to as proportion.

 ***The Three Major forms of ratios are:***

- Ratios that give comparative facts on the component parts in relation to the whole;
- Ratios that relates one observation to another observation, in most cases, observations in the same variables at different time intervals at different places etc.
- Ratio that can be used to compare variables which are not expressed in the same unit.

4.3.2 Percentage

A percentage is a rate, number, or amount in each hundred.

Percentage: $18:36 = \frac{18}{36} X \frac{100}{1}\% = 50\%$

The advantage of percentage is that it makes ratios more comprehensive as it gives a common denominator, 100, to the ratios

4.3.3 Random Numbers

A random number is a number chosen chance from some specified distribution such that selection of a large set of these numbers, reproduce the underlying distribution. Almost always, such numbers are required to be independent, so that there are no correlations between successive numbers.

4.3.3.1 A Random Number Table: A series of digits (0 to 9) arranged **randomly** through the rows and columns.

Recall:

A population is a collection of units with identical properties. The number of units making up the population size is usually denoted by N.

A Sample is a finite number of elements drawn from the population. The sample size is denoted by n. Population and Sample are very important in the use of table of random numbers.

4.3.2.2 Use of Table of Random Numbers

A table of random numbers contain outcomes of independent random trials from the discrete uniform probability distribution which has possible outcomes: 0, 1, 2.....9 with equal probabilities. Thus for each position in the table, every digit from 0-9 has equal probability of appearing in that position and the outcome for the various positions in

the table are independent. A table of random digits can readily be used to select a simple random sample by the following procedure: Consider the frame of lawyers who are members of state Bar Association as of last month, all together they are also in the frame, we assign numbers to the lawyers for convenience from 001,.....950. Since the frame contains 950 elements, we select 3 digits numbers and starting from the left most column, suppose we use the 1st 3 digits in each line as a 3 digits number, the procedure is then as in the next section.

4.3.2.3 Procedure for the use of Table of Random Numbers:

Draw a sample number of size n=20 from a population of 950. Since the population is of 3 digits, 001, 002. 003.....950, Lawyer number 231 is the 1st element for the sample.

- A number over 950 would be disregarded, thus each of the 950 lawyers have equal probability of being a sample element. Lawyer 005 is of second element for the sample.
- Also number 231 would be disregarded since that lawyer is already in the sample, thus each of the 947 has an equal probability of being of 3rd sample element.
- The above procedure is repeated until the required number of lawyers for the sample is selected.

The 20 lawyers are:
231,055,455,070,003,949,555,647,777,090,047,777,701,694,256,901,162,295,007,706.

4.3.3.4 Advantages and Disadvantages of the use of Random Numbers

It is easier to obtain a table of random numbers from a given population but could be time wasting.

SELF-ASSESSMENT EXERCISE(S)

1. If the total expenditure is as given in Table 3.2 and N17 was spent on food, write the ratio of food to total expenditure. Also calculate the ratios and percentages of each of the items in Table 7.2;

Table 3.2: Breakdown of items and cost of each

Item	Expenditure N
Dress	12.00
Shoe	10.00
Watch	9.00
Others	5.00

2. What are three major forms of ratios?
3. Explain how to use the Table of Random Numbers.

4. Describe the Procedure for the use of Table of Random Numbers.
5. What are the advantages and disadvantage of the use of Random Numbers?

4.4 Summary

Some calculations that can be carried out after tabulations of data are usually division of a component of the table by another component of the total. Some other commonly used comparative facts include ratios and percentages**.**

This unit gives a comparative fact on the component parts in relation to the whole and also relates observations in the same variables at different time intervals at different places etc.

4.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Buglear, J. (2002). Stats Means Business: A Guide to Business Statistics, Butterworth-Heinemann, Oxford, ISBN 0-7506-5364-7.

Dodge, Y. (2008). The Concise Encyclopedia of Statistics. Springer. Everitt, B. S.; Skrondal, A. (2010). The Cambridge Dictionary of Statistics, Cambridge University Press.

Gonick, L. (1993). The Cartoon Guide to Statistics. HarperPerennial.

MODULE 3 MEASURES OF CENTRAL TENDENCY, LOCATION AND DISPERSION

Here, location and spread of a distribution of data in units 1, 2 and 3 are discussed.

Unit 1 Measure of Central Tendency/ Location
Unit 2 Fractiles - Measures of Partition and Dispersion
Unit 3 Measures of Dispersion/Spread

UNIT 1 MEASURE OF CENTRAL TENDENCY/ LOCATION

Unit Structure

1.1 Introduction

These are useful in describing statistical data and comparing one group of data with another. The most commonly known measures of central tendency are the arithmetic mean, median, mode, quartiles, deciles, percentiles, harmonic mean, and geometric mean. In all, we estimate the values of measure of central tendency.

1.2 Intended Learning Outcomes (ILOs)

To find a typical or central value that best describes the data.

1.3 Main Content

1.3.1 Measures of Location

- Arithmetic Mean:
- Mode
- Median

1.3.2 Arithmetic Mean:

Arithmetic mean may be defined as the sum of all the values of the items divided by the number of items.
The population mean (theoretical mean) is represented by μ (meu) which is given by $\mu = \frac{\sum x_i}{N}$ 3.1

If a sample of size (n) is taken from the population then the sample mean is represented by $\bar{X} = \frac{\sum x_i}{n}$ 3.2

1.3.3 Basic Arithmetic Mean

Three basic Arithmetic Mean are as follows;

Definition I: Given a set of numbers. X_1, X_2.....X_N, then the mean is given by:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \quad 3.3$$

The mean of the numbers**:** 1, 2, 3, 4, and 5. $= \frac{15}{5} = 3$

Definition II: If the numbers. X_1, X_2.....X_k occur with associated frequencies f_1, f_2.....f_k then $\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i}$ 3.4

Definition III: If X_i is the ith class mark having frequency f_i and k classes, then the mean is obtained by using 3.4.
However, if A is the assumed mean and C the class size of frequency distribution then the mean using the coding methods is given by:

$$\bar{X} = A + \frac{\sum_{i-1}^{k} f_i U_i}{\sum_{i-1}^{k} f_i} XC \quad where\ \mu_i = \frac{X - A}{C} \qquad 3.5$$

1.3.4 Mean of Group Data

Suppose that a given set of data are grouped such that each item x_i (where $i = 1, 2, 3, ...n$) occurs with frequency f_i (where $i = 1, 2, ..., k$) then the mean

$$\bar{X} = \frac{\sum f_i x_i}{\sum f_i} \qquad 3.6$$

1.3.5 Computation of Mean Using Change of Origin

i. **Assumed Mean**: A lot of effort in computation can be saved by reducing each number if they are very large. This is done by selecting a reference number called the assumed mean and then subtract it from other number in the distribution for ungrouped data;

If A=assumed mean, and V=Deviation of each value from the assumed mean

i.e. $V_i = x_i - A\ then$

$$\bar{X} = A + \frac{\sum (x_i - A)}{n} \qquad 3.7$$

$$\bar{X} = A + \frac{\sum V_i}{n} \qquad 3.8$$

For a set of grouped data, the procedure for computing the mean is:

1. Choose the assumed mean (A) which may be one of the class marks
2. Find the deviation $x_i - A$ of each class mark X_i from the assumed mean.
3. Find the product (f_iV_i) and the sum ($\sum f_i V$)

4. Obtain the correction factor

$$\left(\frac{\sum f_i V_i}{\sum f_i}\right) \qquad 3.9$$

5. Add the correction factor to the mean in order to obtain the exact mean for the distribution using the formula:

$$A + \frac{\sum f_i (x_i - A)}{\sum f_i} \qquad 3.10$$

$$\bar{X} = A + \frac{\sum f_i V_i}{\sum f_i} \qquad 3.11$$

ii. Coding Method: If the figures are persistently large, after applying the assumed mean techniques, the computation can further be reduced by dividing each (V_i) by the class width or by the multiples of it. In such a case, we will be using the transformation such that:

$$\mu_i = \frac{x_i - A}{w} \qquad 3.12$$

$$\bar{x} = A + \sum \qquad 3.13$$

and the mean will be defined as

$$\bar{X} = A + \frac{\sum f_i U_i w}{\sum f_i} \qquad 3.14$$

where U_i =0, +1or-1,+2 or-2, if the middle class mark is taken as the assumed mean

1.3.6 Mode

The mode for an ungrouped set of numbers is that value which occurs with the highest frequency. The mode when it exists may always be unique.

For a set of grouped data, the modal class (which is the class with the highest mode) can be clearly ascertained, but the value of mode is not so obvious, but can be calculated by:

i. Interpolation within the modal class.

ii. Graphic interpolation from the histogram

1.3.5 Interpolation within the Modal Class

The mode is located within the modal class by considering the frequency for the two out joining classes to the modal class

$$X_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \text{xw} \qquad 3.16$$

Where L = Lower boundary of the modal class.
Δ_1 = Absolute diff. of the freq. of the modal class and the previous class.
Δ_2 = Diff. between the frequencies of the modal class and that of the lower class.
w = Width of the modal class.

1.4 Median

1.4.1 Median for Grouped Data

The median of a set of numbers arranged in order of magnitude (in array) is the central value (if the set is odd) or the two middle values (if the set is even). The location is at the $\frac{(N+1)}{2}$ item after arrangement.

1.4.2 Median for Ungrouped Data

For a set of ungrouped data, the median is the value in the middle or the mean of the two middle values (depending whether the number of values is odd or even) when the values are arranged in their order of magnitude. In other words, median is that value which divides the distribution into two equal parts.
Symbolically, when the sample size (n) is an odd numbers

$$X_m = \frac{n+1}{2} \qquad 3.15$$

When the sample size (n) is an even numbers

$$X_n = \frac{1}{2}\left(\frac{X_n}{2} + \frac{X_n}{2} + 1\right), \text{ when n is even} \qquad 3.16$$

1.4.3 For a set of ungrouped data, the median can be obtained by the following techniques:

i. Approximation from the frequency distribution.
ii. Interpolation from the Ogive

i. Approximation from the Frequency Distribution

The median being the value of $\frac{N}{2}th$ term has to be in a particular class called the median class with which it can be estimated using the formulae:

$$X_m = L + \frac{\frac{N}{2} - C\ XW}{f} \qquad 3.17$$

where L= Lower boundary of the median class

$\frac{N}{2}$ = median term

C = Cumulative frequency of the class proceeding the median class

F = Freq. of the median class

W = Width of the median class

F = frequency of the median class

ii. Interpolation from the Ogive

When estimating the median from the ogive, the following steps should be followed:

i. Prepare or construct the ogive.

ii. Calculate or locate the $\frac{N}{2}$, the median term on the vertical axis.

iii. Starting from and locate in (ii) draw a horizontal line to intercept the ogive.

iv. Draw a line from the point of intersection to the horizontal and read up the value of the median at the point.

1.4.4 For grouped data the median is obtained by using

$$\text{Median} = L_m + \frac{\frac{N}{2} - F_{m-1}}{F_m} X\ C \qquad 3.18$$

The following formula

$$\overline{\overline{X}} = L_1 + \frac{\left[\frac{N}{2} - (\sum f_{m-1})_1\right]}{f_m} C \qquad 3.19$$

L_1 = lower class boundary of the median class

N = Total frequency

Fm = frequency of the median class

$\frac{N}{2}$ =Location of the median class

C = Size of the median class

$(\sum f_i)_1$ = Sum of frequencies of all classes below the median class.

F_{m-1} = The cumulative frequency of the pre-median class

1.5 Weighted Mean (X_w)

When each value in a given set of data is assigned a weight according to its relative importance in the group, the computed mean is its weighted mean.

An example of a weighted mean is the grade point average of the semester exam or cumulative.

Grade point A of the students (G.P.A). Mean score $\bar{X}_W = \frac{\sum W_i X_i}{\sum W_i}$

3.20

$$X_{GPA} = \frac{\sum X_i G_i}{\sum W_i} \qquad 3.21$$

1.6 Geometric Mean

The geometric mean of a set of ungrouped data $x_1, x_2, ..., x_n$ is nth root of the product of n value involved: $\bar{X}_9 = \sqrt[n]{x_1 \text{x } x_2 \text{ x } x_3 \ldots\ldots\ldots \text{x } x_n} = (x_1 \text{ x } x_2 \ldots\ldots\ldots x_n)^{\frac{1}{n}}$ 3.22

When log is used in computation, the formula below is used.

$$\text{Log } \bar{X}_9 = \frac{1}{n}\sum \log x_i \qquad 3.23$$

$$\bar{X}_G \ Anti\log \left\{\frac{1}{n}\sum \log X_i\right\} \qquad 3.24$$

$$\Rightarrow n = \sum f_i \qquad 3.25$$

For a set of grouped data the geometric mean is defined as

$$Log\ \bar{X}_G = \frac{\sum f_i \log x_i}{\sum f_i} \qquad 3.26$$

$$\bar{X}_G \quad Anti\log \left\{\frac{\sum f_i \log x_i}{\sum f_i}\right\} \qquad 3.27$$

1.6.1 Harmonic Mean

For a set of ungrouped data $x_1, x_2, x_3 \ldots\ldots\ldots x_n$ of a random variable X_1 the harmonic mean is given as:

$$X_H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}\ldots\ldots\ldots\ldots+\frac{1}{x_n}} \qquad 3.28$$

$$\Rightarrow X_H = \frac{\sum f_i}{\sum(\frac{1}{x_i})} \qquad 3.29$$

For a set of grouped data, the harmonic mean formula becomes:

$$X_H = \frac{\sum f_i}{\sum\left(\frac{f_i}{x_i}\right)} \qquad 3.30$$

Self-Assessment Exercise(s)

1. What is the difference between mean, mode and median?
2. What are the three basic arithmetic means?
3. When do we use assumed mean and coding method?
4. Write the median formula for grouped data and explain the parameters.
5. Differentiate between the geometric mean, harmonic mean and weighted mean.

1.7 Summary

It summarizes a list of numbers by a "typical" value. A fundamental task in many statistical analyses is to estimate ***location of*** parameter for the distribution. The three most common **measures of location** are the mean, the median, and the mode.

In this unit, useful in describing statistical data and comparing one group of data with another were discussed, such as the arithmetic mean, median, mode, quartiles, deciles, percentiles, harmonic mean, and geometric mean.

1.8 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Buglear, J. (2002). Stats Means Business: A Guide to Business Statistics, Butterworth-Heinemann, Oxford, ISBN 0-7506-5364-7.

Dodge, Y. (2008). The Concise Encyclopedia of Statistics. Springer. Everitt, B. S.; Skrondal, A. (2010). The Cambridge Dictionary of Statistics, Cambridge University Press.

Gonick, L. (1993). The Cartoon Guide to Statistics. Harper Perennial.

UNIT 2 FRACTILES - MEASURES OF PARTITION AND DISPERSION

Unit Structure

2.1 Introduction

Definition of Fractiles
Fractiles are the measures of partition and the three measures of partition are the Quartile, Decile and the Percentile.

2.2 Intended Learning Outcomes (ILOs)

To examine the different measures of partitions: Quartile, Decile and Percentile

2.3 Main Content

2.3.1 Fractiles can be divided into different parts as follows;

i. First or lower quartile (Q_1) is the value below which 25% of data can be found.
ii. Second quartile (Q_2) is the value below which 50% of the data lie
iii. Third quartile upper quartile (Q_3) is the value below which 75% of the data can be found.

2.3.2 Quartiles

These measures are obtained by adjusting the median formula into four equal parts to locate the appropriate quartile. The quartile is obtained by dividing the set of data into four equal parts.

Thus, yielding first quartile Q_1, 2nd quartile = Q_2 , 3rd quartile = Q_3 and 4th quartile = Q_4.

The component formula for Q_i is

$$Q_i = LQ_i + \frac{\left[\frac{iXN}{4} - \left(\sum fQ_i\right)_1\right]}{fQ_i} XC \qquad 3.1$$

Where Q_i=The ith quartile
fQ_i=Freq. of the ith quartile class
C=Size of the ith quartile class
LQ_i=Lower class boundary of the ith quartile class
$\frac{iXN}{4}$=Location of the cumulative freq. of the ith quartile class
$\left(\sum fQ_i\right)$=Sum of all frequencies lower than the ith quartile class.

2.3.3 Percentiles

The percentiles divide the data into one hundred parts and we have P_1,P_2.....P_{99}.
The percentile class is located where we have the cumulative frequency of

$$\frac{KXN}{100}, \qquad 3.2$$

2.3.4 Deciles

The decile divide the set into ten parts. It then have D_1,D_2,.....D_9. Location of the ith decile is at the point where the cumulative frequency is $\frac{iXN}{10}$
The formula then becomes

$$D_j = LD_j + \frac{\left[\frac{jXN}{10} - \left(\sum fD_j\right)_1\right]}{fD_j} XC \qquad 3.3$$

- **Moments are set of statistical parameters to measure a distribution. Four moments are commonly used:**

i. The first moment as the Mean which is also known as the average
ii. The second moment (about mean?) as the Variance and it is pertinent to note that Standard deviation is the square root of the variance: an indication of how closely the values are spread about the mean. A small standard deviation means the values are all similar. If the distribution is normal, 63% of the values will be within 1 standard deviation.

- **Skewness measures the asymmetry of a distribution about its peak; it is a number that describes the shape of the distribution**.

- It is often approximated by Skew = (Mean - Median) / (Std dev).
- If skewness is positive, the mean is bigger than the median and the distribution has a large tail of high values.
- If skewness is negative, the mean is smaller than the median and the distribution has a large tail of small values.

- **Kurtosis measures the peakedness or flatness of a distribution.**

- Positive kurtosis indicates a thin pointed distribution.
- Negative kurtosis indicates a broad flat distribution.

There are usually three types of kurtosis namely, leptokurtic, platykurtic and mesokurtic.

Self-Assessment Exercise(s)

1. What is fractiles?
2. Explain the difference between first or lower quartile, Second quartile, third quartile and upper quartile.
3. What do you understand by quartiles, percentiles, moment
4. Write the difference between skewness, Kurtosis and deciles.
5. Write the formula for quartiles, percentiles, deciles moment, skewness, Kurtosis and explain their parameters

2.4 Summary

Fractiles are measures of partition for the distribution. The three furthermost common **measures were considered, namely;** Quartile,

Deciles and the Percentile. We also have the moment, skewness and Kurtosis that measures the spread in statistical distributions.

In summary,

- Quartile: divides the data into 4 equal parts;
- Deciles: divides the data into 10 equal parts;
- Percentile: divides the data into 100 equal parts.

2.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

Bhattacharyya, G. K., and R. A. Johnson, (1997). Statistical Concepts and Methods, John Wiley and Sons, New York.

Moore, David S., "The Basic Practice of Statistics." Third edition. W.H. Freeman and Company. New York. 2003.

UNIT 3 MEASURES OF DISPERSION/SPREAD

Unit Structure

3.1 Introduction

The simplest meaning that can be attached to the word ‘dispersion’ is a lack of uniformity in the sizes or quantities of the items of a group or series. According to Reiglemen, “Dispersion is the extent to which the magnitudes or quantities of the items differ; the degree of diversity.” The word dispersion may also be used to indicate the spread of the data.

3.2 Intended Learning Outcomes (ILOs)

To find the basic property of dispersion as a value that indicates the extent to which all other values are dispersed about the central value in a particular distribution.

3.3 Main Content

3.3.1 Definition of Dispersion

Dispersion in statistics is also called variability, scatter, or spread. This is the extent to which a distribution is stretched or squeezed. Common examples of measures of statistical dispersion are the variance, standard deviation, and interquartile range. Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions.

3.3.2 Properties of a good measure of Dispersion

The pre-requisites for a good measure of dispersion:

- It should be simple to understand.
- It should be easy to compute.
- It should be rigidly defined.
- It should be based on each individual item of the distribution.
- It should be capable of further algebraic treatment
- .It should have sampling stability.
- It should not be unduly affected by the extreme items.

3.3.3 Types of Dispersion

The measures of dispersion can either be '**absolute' or "relative".**

3.3.3.1 Absolute measures

Absolute measures of dispersion are expressed in the same units in which the original data are expressed.

For example, if the series is expressed as marks of the students in a particular subject; the absolute dispersion will provide the value in marks. The only difficulty is that if two or more series are expressed in different units, the series cannot be compared on the basis of dispersion.

3.3.3.2 Relative

Relative or Coefficient of dispersion is the ratio or the percentage of a measure of absolute dispersion to an appropriate average. The basic advantage of this measure is that two or more series can be compared with each other despite the fact they are expressed in different units. Theoretically, 'Absolute measure' of dispersion is better. But from a practical point of view, relative or coefficient of dispersion is considered better as it is used to make comparison between series.

3.3.4 Methods of Dispersion

Methods of studying dispersion are divided into two types:
(i) Mathematical Methods: We can study the 'degree' and 'extent' of variation by these methods. In this category, commonly used measures of dispersion are :

- Range

- Quartile Deviation
- Average Deviation
- Standard deviation and coefficient of variation.

(ii) Graphic Methods: Where we want to study only the extent of variation, whether it is higher or lesser a Lorenz-curve is used.

3.3.5 Mathematical Methods

- **Range (**Absolute Range)**:**
  It is the simplest method of studying dispersion. Range is the difference between the smallest value and the largest value of a series. While computing range, we do not take into account frequencies of different groups.
  Formula: Absolute Range = Largest – Smallest = $L - S$ 3.1

- **Coefficient of Range:**
  This is a relative measure of dispersion and is based on the value of the range. It is also called range coefficient of dispersion. It is defined as:

$$\textbf{Coefficient of Range} = \frac{L-S}{L+S} \quad 3.2$$

where, L represents largest value in a distribution
S represents smallest value in a distribution
We can understand the computation of range with the help of examples of different series,

Case Studies_on raw data, discrete series and continuous series
Raw Data: Marks out of 50 in a subject of 10 students, in a class are given as follows: 2,8,2,1,6,4,3,3,8,2,2and 5.

In the example, the maximum or the highest marks is '8' and the lowest marks is '1'. Therefore, we can calculate range;
L = 8 and S = 1
Absolute Range = L – S = 7 marks
Discrete Series

Table 3.1: Marks of the Students in Statistics

Marks of the Students in Statistics (X)	**No. of students (F)**
Smallest 10	4
10	10
18	16
Largest 20	15
	Total = 45

Absolute Range = 20 – 10 = 10 marks

Continuous Series

Table 3.2: Frequencies

Marks of the Students in Statistics (X)	**No. of students (F)**
Smallest= 10 10-15	4
15-20	10
20-25	26
Largest =30 25-30	8
Absolute Range = L – S = 30 – 10 = 20 marks	

Absolute Range = L – S = 30 – 10 = 20 marks

Range is a simplest method of studying dispersion. It takes lesser time to compute the 'absolute' and 'relative' ranges. Range does not take into account all the values of a series, i.e. it considers only the extreme items and middle items are not given any importance. Therefore, Range cannot tell us anything about the character of the distribution. Range cannot be computed in the case of "open ends' distribution i.e., a distribution where the lower limit of the first group and upper limit of the higher group is not given.

The concept of range is useful in the field of quality control and to study the variations in the prices of the shares etc.

 ***Quartile Deviations (Q.D.)***

The concept of 'Quartile Deviation does take into account only the values of the 'Upper quartile (Q3) and the 'Lower quartile' (Q1). Quartile Deviation is also called 'inter-quartile range'. It is a better method when we are interested in knowing the range within which certain proportion of the items fall.

'Quartile Deviation' can be obtained as:

- Inter-quartile range = $Q3 - Q1$ 3.3
- Semi-quartile range = $\frac{Q3-Q1}{2}$ 3.4
- Coefficient of Quartile Deviation $= \frac{Q3-Q1}{Q3+Q1}$ 3.5

 **Case Studies**

Calculation of Inter-quartile Range, Semi-quartile Range and Coefficient of Quartile Deviation in case of Raw Data
Suppose the values of X are : 20, 10, 18, 25, 32, 10
In case of quartile-deviation, it is necessary to calculate the values of Q1

and Q3 by arranging the given data in ascending of descending order. Therefore, the arranged data are (in ascending order): X = 10, 10, 18, 20, 25, 32.

Number of items = 6

Q1 = the value of item = 1.75th item
= the value of 1st item + 0.75 (value of 2nd item – value of 1st item)
= 10 + 0.75 (10 – 10) = 10 + 0.75(2) = 10 + 1.50 = 10.50
Q3= the value of item = the value of 3(7/4)th item = the value of 5.25th item
= 25 + 0.25 (32 – 25) = 25 + 0.25 (7) = 26.075

Therefore,
Inter-quartile range = Q3 – Q1 = 26.75 – 10.50 = 16.25
Semi-quartile range = $\frac{Q3-Q1}{2}$ = 8.125
Coefficient of Quartile Deviation $= \frac{Q3-Q1}{Q3+Q1}$
More examples on Coefficient of Quartile Deviation = $(Q_3 – Q_1) / (Q_3 + Q_4)$
Coefficient of Quartile Deviation = (94.5 – 57.25) / (94.5 +57.25)
Coefficient of Quartile Deviation = 0.2454.
Coefficient of Quartile Deviation = $(Q_3 – Q_1) / (Q_3 + Q)$. Coefficient of Quartile Deviation = (31.25 – 13.50) /(31.25 + 13.50)
Coefficient of Quartile Deviation = 0. 397

Advantages of Quartile Deviation
Some of the important advantages are :
- It is easy to calculate. We are required simply to find the values of Q1 and Q3 and then apply the formula of absolute and coefficient of quartile deviation.
- It has better results than range method. While calculating range, we consider only the extreme values that make dispersion erratic, in the case of quartile deviation, we take into account middle 50% items.
- The quartile deviation is not affected by the extreme items.

Disadvantages
- It is completely dependent on the central items. If these values are irregular and abnormal the result is bound to be affected.
- All the items of the frequency distribution are not given equal importance in finding the values of Q1 and Q3.
- It does not take into account all the items of the series and hence considered to be inaccurate.

Similarly, sometimes we calculate percentile range, say, 90th and 10th percentile as it gives slightly better measure of dispersion in certain

cases.

- Absolute percentile range = P90 – P10.
- Coefficient of percentile range =

This method of calculating dispersion can be applied generally in case of open end series where the importance of extreme values are not considered.

 **Average Deviation**

Average deviation is defined as a value which is obtained by taking the average of the deviations of various items from a measure of central tendency Mean or Median or Mode, ignoring negative signs. Generally, the measure of central tendency from which the deviations arc taken, is specified in the problem. If nothing is mentioned regarding the measure of central tendency specified then deviations are taken from median because the sum of the deviations (after ignoring negative signs) is minimum.

Advantages of Average Deviations

- Average deviation takes into account all the items of a series and hence, it provides sufficiently representative results
- It simplifies calculations since all signs of the deviations are taken as positive.
- Average Deviation may be calculated either by taking deviations from Mean or Median or Mode.
- Average Deviation is not affected by extreme items
- .It is easy to calculate and understand.
- Average deviation is used to make healthy comparisons.

Disadvantages of Average Deviations

- It is illogical and mathematically unsound to assume all negative signs as positive signs.
- Because the method is not mathematically sound, the results obtained by this method are not reliable.
- This method is unsuitable for making comparisons either of the series or structure of the series.

This method is however more effective during the reports presented to the general public or to groups who are not familiar with statistical methods.

Standard Deviation

The standard deviation, which is shown by Greek letter s (read as sigma) is extremely useful in judging the representativeness of the mean. The concept of standard deviation, which was introduced by Karl Pearson has a practical significance because it is free from all defects, which exists in a range, quartile deviation or average deviation. Standard deviation is calculated as the square root of average of squared deviations taken from actual mean. It is also called root mean square deviation. The square of standard deviation i.e., s^2 is called 'variance'.

 ***Calculation of standard deviation in case of raw data***

There are four ways of calculating standard deviation for raw data:

- When actual values are considered;
- When deviations are taken from actual mean;
- When deviations are taken from assumed mean; and
- When 'step deviations' are taken from assumed mean.

Advantages of Standard Deviation

- Standard deviation is the best measure of dispersion because it takes into account all the items and is capable of future algebraic treatment and statistical analysis.
- It is possible to calculate standard deviation for two or more series.
- This measure is most suitable for making comparisons among two or more series about variability.

Disadvantages

- It is difficult to compute.
- It assigns more weights to extreme items and less weights to items that are nearer to mean.
- It is because of this fact that the squares of the deviations which are large in size would be proportionately greater than the squares of those deviations which are comparatively small.

 **Mathematical properties of standard deviation (σ)**

- If deviations of given items are taken from arithmetic mean and squared then the sum of squared deviation should be minimum, i.e., = Minimum,
- If different values are increased or decreased by a constant, the standard deviation will remain the same.
- If different values arc multiplied or divided by a constant than the standard deviation will be multiplied or divided by that constant.

- Combined standard deviation can be obtained for two or more series

Self-Assessment Exercise(s)

1. Write the seven properties of a good measure of Dispersion.
2. Explain the two types of Dispersion.
3. List Methods of Dispersion.
4. Define the following;

a) Dispersion
b) Range
c) Quartile Deviation
d) Average Deviation
e) Standard deviation and coefficient of variation.

5. Write three advantages and two disadvantages of Standard Deviation

3.4 Summary

The concept of central tendency, which is identifying a single value that can be used to represent an entire data set, and also where a distribution of data lies and how the variability or differences among scores in a set of data influences the shape and spread of the distribution were considered in unit 4'

This unit handles the basic property of dispersion as a value that indicates the extent to which all other values are dispersed about the central value in a particular distribution.

3.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Alreck, P.L. and Settle, R. (2003). Survey Research Handbook, 3rd edition, McGraw Hill, New York.

Barnes, S. (Ed.) (2005). News of the World, Football Annual 2005–2006, Invincible Press, London, ISBN 0-00-720582-1.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

MODULE 4 STATISTICAL PROBABILITY (SET THEORY AND BASIC CONCEPTS OF PROBABILITY)

UNIT 1 SET THEORY

Unit Structure

1.1 Introduction

Probability is the extent to which an event is likely to occur. It is the study of random or non-deterministic experiment. Example: Toss of a coin or throw a die in the air.

1.2 Intended Learning Outcomes (ILOs)

To measure by the ratio of the favorable cases to the whole number of cases possible.

1.3 Main Content

1.3.1 Set

A set is a collection of well-defined objects. It is an unordered collection of distinct elements of the same type where type is defined by the writer of the set. A set can also be defined as any collection of objects with definite property which makes them to be viewed as a single entity e.g. students, tribes, teams, electorates. The individual objects in the collection are called elements or members of the set, and they are said to belong to or to be contained in the set. The set in turn is said to contain or be composed of its members. Every element of the set must have the property of the set.

Generally, a set is denoted by a capital symbol while elements of a set are represented by small letters. Members of a set are enclosed in a braced bracket { }. For example, sets are usually denoted by capital letters A, B, C... elements are designated by lowercase letters a, b, c, ... Special notations are used in sets like; $x \in A$ to mean "x is an element of A" or x belongs to A.

A set may be defined by listing the members in brace brackets e.g. the set of positive even numbers less than 10 is denoted by the symbol {2, 4, 6, 8}. The set of all positive even integers is denoted by {2, 4, 6.....}.

1.3.2 Methods of Representing a Set

By Listing: The method of listing the members of a set within braces is sometimes referred to as the **roster notation,** example, A= {1,2,3,4,5, 6}.

By Describing: A set can also be defined by describing the members or the property of the set in brace brackets too;

For example, to write that a set A = {1, 2, 3, 4, 5, 6} in words is A = {Set of numbers gotten consists of all point when a die is rolled}, another example is A = {set of square numbers from 1 to 20}.

By Builder Notation: Another method of describing set mathematically is by using a variable followed by a colon followed by the property of the set, for example, the set of integers between 1 and 10 exclusive is written as {x : 1< x < 10, x is integer}. This is read as the set of all x such that x lies between 1 and 10 exclusive. Also the signs of inequality can be used, example. A= $\{x : 0 \leq x \leq 6\}$

1.3.3 Types of set

There are many types of set in the set theory:

- **Singleton Set or Unit Set**
  A set containing only one element is called a unit set. It is also called a singleton set. Hence the set given by **{1}, {0}, {a}** are all consisting of only one element and therefore are singleton sets.

- **Finite Set (Countable Sets)**
  A set consisting of a finite number of objects or elements in which the number element is countable is said to be a finite set. It can also be a finite set whose members can be countable. Consider the sets: **A = {5, 7, 9, 11} and B = {4, 8, 16, 32, 64, 118}**
  Obviously, **A**, **B** contain a finite number of elements, that is, 4 objects in **A** and 6 in **B**. Thus they are finite sets; other examples are tossing a die, days of the week and so on.

- **Countable Infinite Sets**
  A set that can be countable but has unlimited number of objects examples odd number, even number, that is, y= {2,4,6,8.....}

- **Infinite and Uncountable Sets rational numbers**
  Unlike finite set, it has no limit to counting, if the number of elements in a set is uncountable, the set is said to be an infinite set.
  Thus the set of all natural number is given by **N = { 1, 2, 3, ...}** is an infinite set. Similarly the set of all rational number between 0 and 1 given by
  A = {x:x Q, 0 <x<1} is an infinite set. Even, number x: x is $(0 \leq x \leq 0.5)$.

- **Equal set**
  Equal set can be defined as when two sets A and B are said to be equal or identical consisting of exactly the same elements irrespective of how many times elements are repeated, and we write

A = B. Thus the set A = {2, 4, 6, 8} equals to B = {2, 2, 4, 6, 6, 8}. Also, A = {a, b, c, d}

B = {a, d, c, a, d, c}, it doesn't matter if the elements are repeated.

- **Null set/ empty set**
  It is possible for a set to contain no elements whatever. Such a set is called an **Empty or Null** set and represented by ϕ. We will consider ϕ to be a subset of every set. At times we may write an empty set as { } with no element in it. Note that {.} is not an

empty set. A null set or an empty set is a valid set with no member. **A = { } / phi cardinality of A is 0.**There are two popular representation either **empty curly braces { }** or a **special symbol ϕ phi**.

- **Subset**

A set A is said to be a subset of set B, whenever every element of A also belongs to B, that is, A $\subset$ B. it can also be said that A is contained in B or that B contains A. A subset **A** is said to be subset of **B** if every elements which belongs to **A** also part of the elements in **B**, example,

A= $\{a,b,c\}$; B= $\{a,b,c,d,e,f\}$(Here, A is a subset and a proper subset of B). Also,

if A={1, 2, 3, 4} and B={1, 2, 3, 4} then A is a subset of B.

- **Super Set**

For any two sets, A and B, A is said to be a subset of B written as A$\subset$ B or B is a **super set** of A written as $B \supset A$ if and only if all elements of A are also elements of B= $\{\}$ *and* ϕ =Empty set.

- **Proper Subset**

A set **A** is said to be a proper subset of **B** if **A** is a subset of **B**, **A** is not equal to **B** or **A** is a subset of **B** but **B** contains at least one element which does not belong to **A**. Assuming, A $\subset$ B and there exists at least one element of B which is not contained in A i.e. B $\not\subset$ A then we say A is a **proper subset** of B and we write A$\subset$ B.

- **Improper Set**

Set **A** is called an improper subset of **B** if and only if **A = B**. Every set is an improper subset of itself. However if we have A $\subset$ B and every element in B is also contained in A, that is, B $\subset$ A, then A is called **an improper set** of B and we write: A $\subseteq$ B . We find out that in this case

A = B.

- **Universal Set**

In all application of set theory, when sets of elements are discussed we can have a larger encompassing set which contains all sets under discussion i.e. all the sets will be subsets of this larger set. An example is i students offering courses A, B, and C , a larger set could be the set of students in the school. Such a set is called the Universal set U and vary from one application to the other. It can also be defined as a set put in consideration for all set present. It can also be defined as any set which is a superset of all sets under consideration.

Let A={1, 2, 3}; B={4, 5, 6, 7, 8, 9}; C={0, 1}then the universal set is U={0, 1, 2, 3, 4, 5, 6, 7, 8, 9}
https://www.includehelp.com/basics/set-theory-and-types-of-set-in-discrete-mathematics.aspx

- **Union of Set**
  Given two or more set A, B, ..., the union of the sets is defined as the set of elements found in either A or B or ..... i.e. the set of all elements that can be found in any of the sets and we write
  $A \cup B \cup \ldots\ldots = \{x : x \in A \text{ or } x \in B \text{ or } x \in \ldots\ldots\}$. Combining element of given sets effectively without repetition.

- **Intersect of Set**
  Given two or more sets A, B... the intersection of the sets is the set of all elements found in A, B and ..., that is, the set of elements belonging to every one of those sets and we write:
  $A \cap B \cap ... = \{x : x \in A, x \in B, x \in \ldots.\}$
  For example, given that A = {1, 2, 3,4, 5}; B = {2, 4, 6, 8,}, then $A \cap B = \{2, 4\}$.
  It involves sorting out common elements in the given sets.

- **Compliment:**
  Given a universal set U, and a set A which is a subset of the universal set, the complement of the set A with respect to U written as A' or A^c is the set of all elements not found in A but are in the universal set, that is, $A' = \{x : x \notin A, x \in U\}$.

- **Disjoint:** When the intersection of two or more sets is empty.

- **Difference of two Sets**:
  Given two sets A, B, the difference A - B (also called the complement of B relative to A) is defined to be the set of all elements of A which are not in B.
  $\therefore \quad A - B = \{x: x \in A, x \notin B\}$.
  Thus we will see that $A - B = A \cup B'$

- **Product Set**
  Show AxB=BxA, where A=$\{1,2\}$; B=$\{a,b\}$ then AxB=(1,a), (1,b), (2,a), (2,b) and
  BxA=(a,1), (b,1), (a,2), (b,2)

- **Partition of Sets**
  This means separating the set with parenthesis that would contain all the elements of the set depending on your objective(s)

Example: partition six even numbers as unit sets from set A; A= {1, 2, 3, 4, 5,…, 13}, we have A= {2}, {4}, {6}, {8}, {10}, {12},

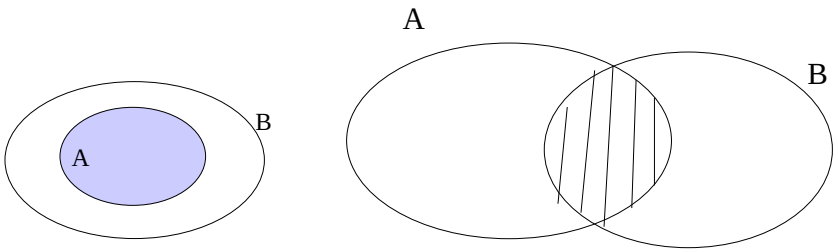
1.3.4 Set Theory

Set theory is a branch of mathematics that deals with the properties of well-defined collections of objects such as numbers or functions. Set theory is also a well-defined collection of definite objects of perception or thought and the Georg Cantor is the father of set theory. A set may also be thought of as grouping together of single objects into a whole. The objects should be distinct from each other and should be distinguished from all those objects that do not form the set under consideration. Hence a set may be a bunch of grapes, a tea set or it may consist of geometrical points or straight lines;
https://www.includehelp.com/basics/set-theory-and-types-of-set-in-discrete-mathematics.aspx

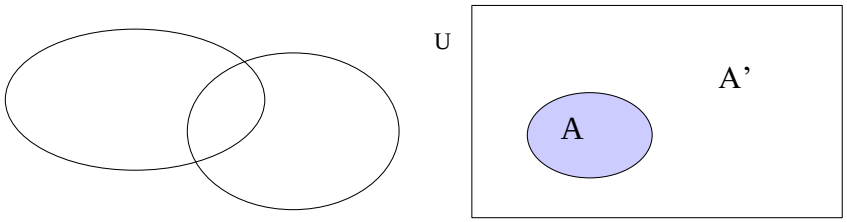
Sets are represented in diagrammatic form by what is called the Venn Diagram

Case Studies

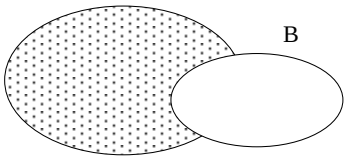
We can demonstrate, the concept of subsets, intersection, union, complement, disjoint, and difference of set in Venn diagrams as follows:



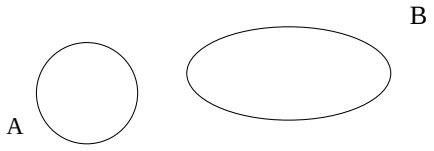
$A \subset B$ $A \cap B$



A∪B



A-B



$A \cap B = \phi$

Number of Elements in a Set:
The number of elements in a set A written n(A) is the number of distinct elements in the set.

Case Study 1, given a set A
A = {1, 2, 1, 2, 2, 3, 3},
There are three distinct elements 1, 2, 3.
∴ n(A) = 3.

Maximum no. of Subsets a Set can have:
If we are given a set A = { }, the subset it has is itself only so the number of subset is 1.
If A= {a}, the subsets are { } and {a}, thus the number of subsets are 2.

Case Study 2

If A = {a, b}, the subsets are { }, {a}, {b}, {a, b}. Therefore the number of subsets is 4. If A = {a, b, c}.
All the subsets of A are:
{a,}, {b,} {c,}, {a, b} {b, c} {a, c}, {a, b, c} and { }
We thus set up a table as follows

No of elements	No of subsets	No of subsets in powers of 2
0	1	2^0
1	2	2^1
2	4	2^2
3	8	2^3
. .	. .	. .
N		2^n

Therefore, for any set having n distinct elements, the maximum number of subsets it can have is 2^n, which is called the power set.

Power set of a set is defined as a set of every possible subset. If the cardinality of **A** is **n** then Cardinality of power set is 2^n as every element has two options either to belong to a subset or not.

Laws of Set Algebra

Given three sets, A, B, C.

(i) Commutative Laws

$A \cup B = B \cup A$ (Commutatively of set).

$A \cap B = B \cap A$

(ii) Associative Laws

$(A \cup B) \cup C = A \cup (B \cup C)$ (Associativity of set)

$(A \cap B) \cap C = A \cap (B \cap C)$

(iii) Distributive Laws

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (Distributive laws).

$A \cap (B \cup C) = (A \cap B) \cup (A \cup C)$

(iv) Idempotent Laws

$A \cup A' = U$ (Idempotent law)

$A \cap A' = \varnothing$

(v) Identity law

$A \cup \Phi = A$ (Identities)

$A \cap U = A$

$A \cup U = U$

$A \cap \Phi = \Phi$

(vi) Complement Law

$A \cup A' = U$

$A \cap A' = \Phi$ (Complement Law)

$(A')' = A$

$U' = \Phi$

$\Phi' = U$

(vii) De Morgan's Laws

$(A \cup B)' = A' \cap B'$ (De Morgan's Laws)

$(A \cap B)' = A' \cup B'$

SELF-ASSESSMENT EXERCISE(S)

1.	Define the following;
a)	Set
b)	Subset
c)	Universal Set
d)	Union Of Set
e)	Intersect of Set
f)	Compliment Set
g)	Product Set
2.	List method of representation a Set and include an example.
3.	List and state mathematically the six laws of sets of algebra
4.	Distinguish between finite, countable infinite sets and infinite and uncountable sets
5.	What do you understand by partition of sets

1.4 Summary

In unit 1, many aspects of Set were discussed, this includes; combining elements of given sets effectively without repetition, sorting out common elements in the given sets, set theory, types of sets and some laws of sets of algebra.

This unit covers the ratio of the favorable cases to the whole number of cases possible on set theory, combination of elements of given sets effectively without repetition, sorting out common elements in the given sets, set theory, types of sets and some laws of sets of algebra.

1.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Alreck, P.L. and Settle, R. (2003). Survey Research Handbook, 3rd edition, McGraw Hill, New York.

Barnes, S. (Ed.) (2005). News of the World, Football Annual 2005–2006, Invincible Press, London, ISBN 0-00-720582-1.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

UNIT 2 BAYES'S THEOREM AND COUNTING TECHNIQUES

Unit Structure

2.1 Introduction

Counting means determining the number of elements in a given set. Counting techniques can be seen as some approaches for determining without direct enumeration the number of all possible outcomes of a conceptual experiment or the number of elements in a given set.

The Fundamental **Counting** Principle (also called the **counting** rule) is a way to figure out the number of outcomes in a probability problem. Basically, you multiply the events together to get the total number of outcomes.

Case Study 1

1. Assuming there are two ways of travelling from Nsukka to Enugu and subsequently three ways of travelling from Enugu to Aba, then there are total of 6 ways of travelling from Nsukka to Aba.
2. If you have 3 shirts and 4 pants; it results to $3\times4=12$ different outfits.
3. If there are 6 flavors of ice-cream, and 3 different cones; there will then be $6\times3=18$ different single-scoop ice-creams you could order

2.1 Intended Learning Outcomes (ILOs)

To ascertain the exact numbers of sub-sets in a given set.

2.3 Main Content

2.3.1 Fundamental Principles of Counting Techniques

- **Multiplicative Rules-**
  If a procedure can be performed in n_1 way, and if following this procedure, a second one can be performed in n_2 ways and so on, then the number of ways the procedure can be performed in the order indicated is the product.

  In combinatorics, the **rule** of product or **multiplication principle** is a basic **counting principle** (a.k.a. the fundamental **principle** of **counting**). Stated simply, it is the idea that if there are a ways of doing something and b ways of doing another thing, then there are a · b ways of performing both actions. It is denoted by $n_1 x n_2 x n_3 x ...$

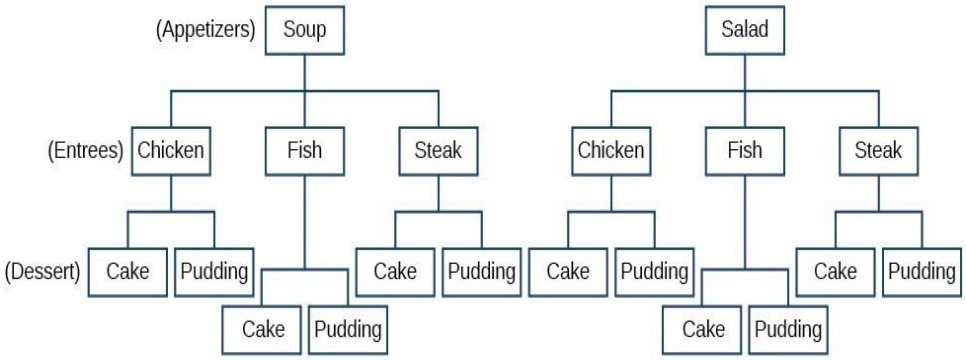
Application of Multiplicative rule

The **Multiplication Rule** of Probability is **used** to find the intersection of two different sets of events, called independent and dependent events. Independent events are when the probability of an event is not affected by a previous event.

Using the Multiplication Principle

Case Study 2

The Multiplication Principle applies when we are making more than one selection. Suppose we are choosing an appetizer, an entrée, and a dessert. If there are 2 appetizer options, 3 entrée options, and 2 dessert options on a fixed-price dinner menu, there are a total of 12 possible choices of one each as shown in the tree diagram below;



The possible choices are:

1. soup, chicken, cake
2. soup, chicken, pudding
3. soup, fish, cake
4. soup, fish, pudding
5. soup, steak, cake
6. soup, steak, pudding
7. salad, chicken, cake
8. salad, chicken, pudding
9. salad, fish, cake
10. salad, fish, pudding
11. salad, steak, cake
12. salad, steak, pudding

We can also find the total number of possible dinners by multiplying. We could also conclude that there are 12 possible dinner choices simply by applying the Multiplication Principle.

# of appetizer options	×	# of entree options	×	# of dessert options	
2	×	3	×	2	= 12

Case Study 3: Using the Multiplication Principle

Diane packed 2 skirts, 4 blouses, and a sweater for her business trip. She will need to choose a skirt and a blouse for each outfit and decide whether to wear the sweater. Use the Multiplication Principle to find the total number of possible outfits.

Solution

To find the total number of outfits, find the product of the number of skirt options, the number of blouse options, and the number of sweater options.

# of skirt options	×	# of blouse options	×	# of sweater options	
2	×	4	×	2	= 16

There are 16 possible outfits.

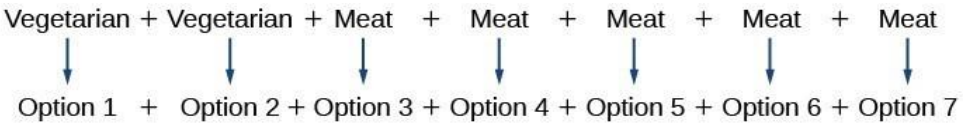
- **Additive Rule**-
  The additive principle states that if event A can occur in m ways, and event B can occur in n disjoint ways, then the event "A or B " can occur in m+n ways. Also, if a procedure can be done in m_1 ways and another done in m_2 ways, then the number of ways the procedure can be done in the other is (m_1+m_2) ways. It is important that the events be disjoint : i.e., that there is no way for A and B to both happen at the same time.

 **Application of Additive rule**

Addition Rule 1: When two events, A and B, are mutually exclusive, the probability that A or B will occur is the sum of the probability of each event.

Addition Rule 2: When two events, A and B, are non-mutually exclusive, there is some overlap between these events.

 Case Studies 4: Using the Addition Principle
If there are 2 vegetarian entrée options and 5 meat entrée options on a dinner menu, what is the total number of entrée options?
SOLUTION
We can add the number of vegetarian options to the number of meat options to find the total number of entrée options.



There are 7 total options.

https://courses.lumenlearning.com/ivytech-collegealgebra/chapter/using-the-addition-and-multiplication-principles/

 In general, the multiplication principle, similar to the addition principle, tells us that if we multiply or divide by a number on one side of an equation, we also need to multiply or divide by that same number on the other side to keep the equation the same.

Subtraction can be an additive relationship because subtracting a number is the same as adding a negative number (example: 5 - 2 = 5 + (-2). Multiplicative relationships mean you multiply any x-value times the SAME number to get the corresponding y-value.

2.3.2 Axioms of Probability

For philosophers, an axiom is a statement like "something can't be true and not be true at the same time." An example of a mathematical axiom is "a number is equal to itself." In everyday usage, an axiom is just a common saying, but it's one that pretty much everyone agrees on.

The axioms of probability are these three conditions on the function P:

- The probability of every event is at least zero.
- The probability of the entire outcome space is 100%.
- If two events are disjoint, the probability that either of the events happens is the sum of the probabilities that each happens.

Unfortunately you can't prove something using nothing. You need at least a few building blocks to start with, and these are called Axioms. Mathematicians assume that axioms are true without being able to prove them. If there are too few axioms, you can prove very little and mathematics would not be very interesting.

 **The axioms of probability can be defined mathematically as**;

P(.) - a generalized symbol of probability which can assume any digit between 0 and 1.
Let P(A) be a real valued function defined on every event $A \in U$..
Then P(A) is called the probability of event A if the following are satisfied:

i) All probabilities are real values between 0 and 1; $0 \leq P(A) \leq 1$ but for $\forall\, A \in U; P(A) \geq 0$
ii) The Valid propositions have probability 1; $P(U) = 1$
iii) The probability of disjunction (mutually exclusive event) is defined as;$P(A \cup B) = P(A) + P(B) - P(A \cap B)$. If $P(A \cap B) = 0$, then

$$P(A \cup B) = P(A) + P(B)$$

2.3.3 Mutually Exclusive

Mutually exclusive is a statistical term describing two or more events that cannot happen simultaneously. It is commonly used to describe a situation where the occurrence of one outcome supersedes the other. Mutually exclusive events are things that can't happen at the same time. For example, you can't run backwards and forwards at the same time. The events “running forward” and “running backwards” are mutually exclusive. Tossing a coin can also give you this type of event.

When a sample space is distributed down into some mutually exclusive events such that their union forms the sample space itself, then such events are called exhaustive events. When two or more events form the sample space collectively then it is known as collectively exhaustive events.

2.3.4 Mutually Exclusive

First, “mutually exclusive” is a concept from probability theory that says two events cannot occur at the same time but at least one of the events must occur. For example, when rolling a six-sided die, the outcomes 1, 2, 3, 4, 5, and 6 One example of an event that is both collectively exhaustive and mutually exclusive is tossing a coin. The outcome must be either heads or tails, or p (heads or tails) = 1, so the outcomes are collectively exhaustive.

When applied to information, mutually exclusive ideas would be distinctly separate and not overlapping. Secondly, “collectively exhaustive” means that the set of ideas is inclusive of all possible options. A collection of events is exhaustive if at least one of them must occur. A collection of events is non-exhaustive if it is possible for none of them to occur. Events are mutually exclusive if no two of them can occur simultaneously.

In probability theory and logic, a set of events is jointly or collectively exhaustive if at least one of the events must occur. ... The set of all possible die rolls is both mutually exclusive and collectively exhaustive (i.e., "MECE"). The events 1 and 6 are mutually exclusive but not collectively exhaustive.

 **Case Studies 5:**

Assuming that, a bag contains 12 white and 8 black balls.
(a) What is the probability of picking a white excludes picking a black, when the events are mutually exclusive;

(b) What is the probability of picking a black excludes picking a white, when the events are mutually exclusive.

White A	Black B	Total	P(A)	P(B)
12	8	20	3/5	2/5

Solution;

a) P(A) = 12/20
b) P(B) = 8/20

2.3.5 Bayes' Theorem

Bayes' Theorem, named after 18th-century British mathematician Thomas **Bayes**, is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome.

 **Application of Bayes Theorem**

It provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence. In finance, **Bayes' theorem** can be **used** to rate the risk of lending money to potential borrowers.

The **Bayes theorem** describes the probability of an event based on the prior knowledge of the conditions that might be related to the event. If we **know** the conditional probability, we can **use** the **Bayes rule** to **find out** the reverse probabilities

Definition of Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad 3.1$$

where;
P(A) is the probability of event A
P(B) is the probability of event B
P(A|B) is the probability of observing event A if B is true
P(B|A) is the probability of observing event B if A is true.

Case Studies 6;

The table 3.1 shows the number of persons who were reported dead in a certain community during 2020 and were classified by sex and age.
Table 3.1: Dead in a certain community during 2021 COVID 19 pandemic

Age/sex	A_1	A_2	A_3	total
	<50yrs	50-79	80+	
Male	27	79	21	128
Female	18	23	20	61
$P\left(A_i/A\right)$	45	102	42	189

If randomly selected individual from the table is male. Find the probability that he is aged between 50-79years using Bayes theorem.

Solution

$$P\left(A_i/A\right)=\frac{P(A_i)P\left(A/A_i\right)}{\sum P(A_i)P\left(A/A_i\right)}=P\left(A_i/A\right)=\frac{P\left(\frac{102}{189}\right)x\left(\frac{79}{102}\right)}{\frac{45}{189}x\frac{27}{45}+\frac{102}{189}x\frac{79}{102}+\frac{42}{189}x\frac{22}{42}}$$

$$=P\left(A_i/A\right)=\frac{\left(\frac{79}{189}\right)}{\frac{27}{189}+\frac{79}{189}+\frac{22}{189}}=\frac{79}{189}X\frac{189}{28}=\frac{79}{28}$$

Case Studies 7: Wiggins's explanation can be summarized with the help of the following table which illustrates the scenario in a hypothetical population of 10,000 people:

	Diseased	**Not Diseased**	
Test +	99	99	198
Test -	1	9,801	9,802
	100	9,900	10,000

In this scenario P(A) is the unconditional probability of the disease; here it is 100/10,000 = 0.01.

P(B) is the unconditional probability of a positive test; here it is 198/10,000 = 0.0198..

What we want to know is P (A | B), i.e., the probability of disease (A), given that the patient has a positive test (B). We know that prevalence of

disease (the unconditional probability of disease) is 1% or 0.01; this is represented by P(A).

Therefore, in a population of 10,000 there will be 100 diseased people and 9,900 non-diseased people. We also know the sensitivity of the test is 99%, i.e., P(B | A) = 0.99; therefore, among the 100 diseased people, 99 will test positive. We also know that the specificity is also 99%, or that there is a 1% error rate in non-diseased people. Therefore, among the 9,900 non-diseased people, 99 will have a positive test. And from these numbers, it follows that the unconditional probability of a positive test is 198/10,000 = 0.0198; this is P(B).

Thus, P(A | B) = (0.99 x 0.01) / 0.0198 = 0.50 = 50%.

From the table above, we can also see that given a positive test (subjects in the Test + row), the probability of disease is 99/198 = 0.05 = 50%.
https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability6.html

Case Studies 8:

Suppose a patient exhibits symptoms that make her physician concerned that she may have a particular disease. The disease is relatively rare in this population, with a prevalence of 0.2% (meaning it affects 2 out of every 1,000 persons). The physician recommends a screening test that costs $250 and requires a blood sample. Before agreeing to the screening test, the patient wants to know what will be learned from the test, specifically she wants to know the probability of disease, given a positive test result, i.e., P(Disease | Screen Positive).

The physician reports that the screening test is widely used and has a reported sensitivity of 85%. In addition, the test comes back positive 8% of the time and negative 92% of the time.

The information that is available is as follows:

- P(Disease)=0.002, i.e., prevalence = 0.002
- P(Screen Positive | Disease)=0.85, i.e., the probability of screening positive, given the presence of disease is 85% (the sensitivity of the test), and
- P(Screen Positive)=0.08, i.e., the probability of screening positive overall is 8% or 0.08. We can now substitute the values into the above equation to compute the desired probability,

Based on the available information, we could piece this together using a hypothetical population of 100,000 people. Given the available

information this test would produce the results summarized in the table 12.2;

Table 12.2: Summary of Positive and Negative Test Results

	Diseased	Not diseased	
Test +	170	7,830	8,000
Test -	30	91,970	92,000
	200	99,800	100,000

The answer to the patient's question also could be computed from Bayes's Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

We know that P(Disease)=0.002, P(Screen Positive | Disease)=0.85 and P(Screen Positive)=0.08. We can now substitute the values into the above equation to compute the desired probability,

P(Disease | Screen Positive) = (0.85)(0.002)/(0.08) = 0.021.

If the patient undergoes the test and it comes back positive, there is a 2.1% chance that he has the disease. Also, note, however, that without the test, there is a 0.2% chance that he has the disease (the prevalence in the population). In view of this, do you think the patient have the screening test?

Another important question that the patient might ask is, what is the chance of a false positive result? Specifically, what is P(Screen Positive| No Disease)? We can compute this conditional probability with the available information using Bayes Theorem.

$$P(\text{Screen positive} \mid \text{No disease}) = \frac{(P(\text{No disease} \mid \text{Screen positive}) \times (P(\text{Screen positive})}{P(\text{No disease})}$$

By substituting the probabilities in this scenario, we get:

$$P(\text{Screen positive} \mid \text{No disease}) = \frac{(1-0.021)(0.08)}{(1-0.002)} = 0.078$$

Thus, using Bayes Theorem, there is a 7.8% probability that the screening test will be positive in patients free of disease, which is the false positive fraction of the test.

Complementary Events

Note that if P(Disease) = 0.002, then P(No Disease)=1-0.002. The events, Disease and No Disease, are called complementary events. The "No Disease" group includes all members of the population not in the "Disease" group. The sum of the probabilities of complementary events must equal 1 (i.e., P(Disease) + P(No Disease) = 1). Similarly, P(No Disease | Screen Positive) + P(Disease | Screen Positive) = 1.

2.3.6 Probability Density Functions

The function f(x) is called the probability density functions of a discrete random variable x if:

(i) $f(x) \geq 0$

(ii) $\sum_{x} f(x) = 1$

Let x be random variable of a continuous type, then f(x) is called a probability density function of the continuous if

(i) $f(x) \geq 0$

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

(iii) $P(a \leq x \leq b) = \int_{a}^{b} f(x)dx$

2.3.7 Cumulative Density Function

As the name implies F(x)-P($X \leq x$)

When discrete, it's expression would be:

(i) if x is discrete; $F(x) = \sum_{u \leq x} f(u)$

(ii) if x is continuous ; $F(x) = \int_{-\infty}^{x} f(x)dx$

Case Studies 9:

The entries of a coin tossed twice

X	0	1	2	Total
F	1	2	1	4
f(x)	¼	2/4	¼	1
f(x)	¼	¾	4/4	2

$$\binom{n}{x}\binom{2}{0} = \frac{2!}{(2-0)!0!} = 1$$

Recall $f(x)=P[X \leq x]$

$F(0)=P[X \leq 0]=1/4$

$F(1)=P[X \leq 1]=P[X \leq 0,1]=1/4+2/4 = \frac{1+2}{4} = \frac{3}{4}$

$F(2) = P[X \leq 2]=P[X \leq 0,1,2]=1/4+2/4+1/4 = \frac{1+2+1}{4} = \frac{4}{4} = 1$

2.3.8 Step Function Graph

Step function is a non-decreasing function, that is, $x_1<x_2$, $(x_1)<f(x_2)$

X	0	1	2	3
F	1	3	3	1
f(x)	1/8	3/8	3/8	1/8
f(x)	1/8	4/8	7/8	8/8

$^3C_0 = 1$; $^3C_1 = 3$; $^3C_3 = 1$

 **Case Study 10:**

Consider f(x)= $\int C(x^2)\ 0 \le x \le 3$

O elsewhere

If f(x) is a pdf

i) find c

ii) p(x<0)

iii) $p(0 \le x \le 1)$

Solution

i) $f(x)=\int_0^3 f(x)dx = \int_0^3 cx^2 dx$

$$= \Rightarrow c\left(\frac{x^3}{3}\right)_0^3 = 1 \rightarrow discrete\ pdf \qquad eq(1)$$

C=1/9

$\int_0^3 \frac{x^2}{9}$

From eq (1) substitute for 1/9= C

$\left(\frac{x^3}{3x9}\right)=\frac{27}{27}=1$ It is a pdf

,

ii) $P(0 \le x \le 1)$

$$\int_0^1 \frac{x}{9}=\left[\frac{x^3}{3x9}\right]_0^1=\left(\frac{1^3}{27}-\frac{0^3}{27}\right)=\frac{1}{27}-0=\frac{1}{27}$$

i. $P(2 \le x \le 3)$

$$\int_2^3 \frac{x^2}{9}=\left[\frac{x^3}{27}\right]_2^3=\left[\frac{3^3}{27}-\frac{2^3}{27}\right] \quad =\left[\frac{27}{27}-\frac{8}{27}\right] = 1-\frac{8}{27} \quad =\frac{27-8}{27} \quad =$$

$\frac{19}{27}$

 **Case Study 11:**

Given, f(x) = $\int 3e^{-3x}, x > 0$, check if it is a proper pdf

Solution

To check if the equation above is a proper pdf;

$\int_1^{\infty} 3e^{-3x} = \left[-9e^{-3x}\right]_1^{\infty} = \left[-9e^{-3(\infty)}\right] - \left[-9e^{-3}\right]$, Not a proper pdf

Using f(x) = $\int 3e^{-3x}, x > 0$

For (i) P $P[x \geq 3]$

$\int_3^{\infty} 3e^{-3x}dx = 3\int_3^{\infty} e^{-3x}dx = 3\left[-\frac{1}{3}e^{-3x}\right]_3^{\infty} = -\left[0 - e^{-9}\right]$=1.23x10^{-4}

SELF-ASSESSMENT EXERCISE(S)

1. What are counting techniques?
2. Discuss the fundamental principles of counting techniques
3. What are the axioms of probability?
4. Two telephones A and B are in use 0.05 and 0.03 of the total time available respectively. A is old and breaks down with probability of 0.2 per unit time while B is new and breaks down with probability of 0.4 per unit if a failure indicator f, shows that one of the telephone is out of order. Find the probability that it is A.
5. Explain the following;
   i. Bayes Theorem
   ii. Probability density functions for discrete and continuous type
   iii. Cumulative density function
   iv. Step function graph

 **2.4 Summary**

Unit 2 determines without direct enumeration, the number of all possible outcome of a conceptual experiment or the number of elements in a given set.

Here the Counting techniques was considered as some approaches for determining without direct enumeration the number of all possible outcomes of a conceptual experiment or the number of elements in a given set.

2.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Alreck, P.L. and Settle, R. (2003). Survey Research Handbook, 3rd edition, McGraw Hill, New York.

Barnes, S. (Ed.) (2005). News of the World, Football Annual 2005–2006, Invincible Press, London, ISBN 0-00-720582-1.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

UNIT 3 PERMUTATIONS AND COMBINATIONS

Unit Structure

3.1 Introduction

In drawing with replacement, each drawing is made from the entire population such that each element can be drawn more than once. The samples so drawn are those samples in which repetitions are allowed.

In drawing without replacement, an element once drawn is removed from the population so that the samples drawn do not allow for repetitions.

3.2 Intended Learning Outcomes (ILOs)

By the end of this unit you are excepted:

- To arrange the objects or numbers in order.
- To ascertain the way of selecting the objects or numbers from a group of objects or collection, in such a way that the order of the objects does not matter.

3.3 Main Content

3.3.1 Permutation

A permutation is an ordered arrangement of a set of n objects. If $r, (r \leq n)$ of these objects are considered it is called an r-permutation or a permutation of n objects taking r at a time.

Case Study 1

If you have a set containing 3 elements; $\{a, b, c\}$ and you are to arrange 2 objects from the 3 objects, in this case, you do not replace the element once selected and in an ordered, then you will have the following $\{a, b\}$, $\{a, c\}$, $\{b, c\}, \{b, a\}$, $\{c, a\}$, $\{c, b\}$results which is equal to 6 permutations. This is called an ordered arrangement without replacement.

i. It is denoted by ${}^{n}P_{r}$ *or* $P(n,r)$ and evaluated in drawing without replacement as ${}^{n}P_{r} = \frac{n!}{(n-r)!} = n(n-2)....(n-r+1)$ 3.1

ii. If arrangement is with replacement then the number of ways is given by n^{r} 3.2

Case Study 2

Assuming, you have a set containing 3 elements; $\{a, b, c\}$ and you are to arrange 2 objects from the 3 objects, in this case, you will replace the element that mean, there is possibility of selecting that particular element more than once and in an ordered, then you will have the following $\{a, b\}$, $\{a, c\}$, $\{b, c\}, \{b, a\}$, $\{c, a\}$, $\{c, b\}$ $\{a, a\}$, $\{b, b\}$, $\{c, c\}$ results which is equal to 9 permutations. This is called an ordered arrangement with replacement.

iii. If we have n objects are arrange taking all n at the same time and if we observe that n_1 are alike, n_2 are alike, and so on then $\sum n_1 = n;$ and the number of arrangements is given by

$$\frac{n!}{n_1!n_2...n_r!}$$ 3.3

Case Study 3

If an urn contains 10 balls, find the number of ordered arrangements

i. Of size 3 with replacement

ii. Of size 5 without replacement

Solution

i. $n = 10$ *and* $r = 3$; with replacement number of ways is n^r .

$$n^r = 10^3 = 1000$$

ii. n=10, r=3 without replacement, the required number of ways is

$${}^{n}P_r = {}^{10}P_5 = \frac{10!}{5!} = 10X9X8X7X6 = 30240$$

Case Study 4

Find the number of permutations that can be formed using the words (i) queue and (ii) statistics

Solution:

i. Queue: $n = 5, e = u = 2, q = 1$

Probability

No of ways $= \frac{n!}{n_1!n_2!n_3!} = \frac{5!}{2!2!1!} = 30$

ii. Statistics $n = 10, t = s = 3, i = 2, a = c = 1$

No of ways: $\frac{n!}{n_1!n_2!n_3!n_4!n_5!} = \frac{10!}{3!X3!X2!X1!X1!} = 50400$

3.3.2 Combination

A combination is an unordered selection of n objects. If $r(r \le n)$ of these objects are selected, we call it an r-combination or a combination of n objects taking r at a time.

i. It is denoted by

$${}^{n}C_r \; or \; \binom{n}{r} \qquad 3.4$$

ii. In sampling without replacement, the evaluation becomes

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \qquad 3.5$$

iii. In sampling with replacement, the number of unordered selection becomes

$$\binom{n+r-1}{r} = \binom{n+r-1}{n-1} = \frac{(n+r-1)!}{r!(n-1)!} \qquad 3.6$$

Case Study 5

Consider the four letters A,B,C,D. How many unordered selections are possible taking 3 letters at a time if

i. Replacement is allowed?
ii. Replacement is not allowed?
iii. List samples in (i) and (ii)

Solution

i. With replacement, total number of ordered samples

$$\binom{n+r-1}{r}, but\ n=4\ and\ r=3;\ therefore \binom{n+r-1}{r}=\binom{6}{3}=20$$

ii. Without replacement

$$\binom{n}{r}=\binom{4}{3}=4$$

The twenty years of listing (i) are:

AAA AAB AAC AAD ABB
ABC ABD ACC ACD ADD
BBB BBC BBD BCC BCD
BDD CCC CCD CDD DDD

The listing of (ii) is: ABC ABD ACD BCD

Case Study 6:

Repeat example 3.3 using 3 letters A, B and C for both (a) ordered and (b) unordered selections of 2 letters out of the three.

Solution

a. Ordered selection – permutation

i. With replacement: total number of samples is

$$n^r = 3^2 = 9.$$

ii. Without replacement: ${}^nP_r = {}^3P_2 = 3.2 = 6$

iii. With replacement the samples are: AB, AC, BC, BA, CA, CB, AA, BB, CC; without replacement: AB, AC, BC, BA, CA, CB.

b. Unordered selection – combination

i. With replacement: the total number of samples is

$$\binom{n+r-1}{n-r}=\binom{4}{2}=\binom{4.3}{2.1}=6$$

ii. Without replacement: total number of sample is

$$\binom{r}{n} = \binom{3}{2} = \binom{3.2}{2} = 3$$

iii. With replacement: the list is AB, AC, BC, AA, BB, CC.
Without replacement: the list is AB, AC, BC.

Case Study 6

If you have a set containing 3 elements; $\{a, b, c\}$ and you are to arrange 2 objects from the 3 objects, in this case, you do not replace the element once selected and in an ordered, then you will have the following; $\{a, b\}$, $\{a, c\}$, $\{b, c\}$, results which is equal to 3 combinations. This is called an ordered arrangement without replacement.

SELF-ASSESSMENT EXERCISE(S)

1. State three differences between Permutation and combination.
2. An urn contains 20 balls, find the number of ordered arrangements
   a. Of size 3 with replacement
   b. Of size without replacement
3. Find the number of permutations that can be formed using the words;
   a. Queue: $n = 5, e = u = 2, q = 1$
   b. Statistics $n = 10, t = s = 3, i = 2, a = c = 1$
4. Consider the four letters A, B, C, D, E. How many unordered selections are possible taking 4 letters at a time if
   a. Replacement is allowed?
   b. Replacement is not allowed?
   c. List samples in (a) and (b)

1.4 Summary

Permutations and combinations are the various ways in which objects from a set may be selected, generally without replacement, to form subsets. This selection of subsets is called a **permutation** when the order of selection is a factor and a **combination** when order is not a factor.

i. In permutation and combination objects being arranged or selected must be distinguishable.
ii. The distinction between combination and permutation is clarified from the distinction between ordered and unordered sampling; arrangements with and without replacement.
iii. The relationship between permutation and combination is that ${}^{n}P_{r} = r!\,{}^{n}C_{r}$.

1.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Alreck, P.L. and Settle, R. (2003). Survey Research Handbook, 3rd edition, McGraw Hill, New York.

Barnes, S. (Ed.) (2005). News of the World, Football Annual 2005–2006, Invincible Press, London, ISBN 0-00-720582-1.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

UNIT 4 BASIC CONCEPTS OF PROBABILITY

Unit Structure

4.1 Introduction

Probabilities can be expressed as proportions that range from 0 to 1, and they can also be expressed as percentages ranging from 0% to 100%. A probability of 0 indicates that there is no chance that a particular event will occur, whereas a probability of 1 indicates that an event is certain to occur.

4.2 Intended Learning Outcomes (ILOs)

To ascertain the number that reflects the chance or likelihood that a particular event will occur.

4.3 Main Content

A probability of 0.45 (45%) indicates that there are 45 chances out of 100 of the event occurring.

The concept of probability can be illustrated in the context of a study of obesity in children 5-10 years of age who are seeking medical care at a particular paediatric practice. The population (sampling frame) includes all children who were seen in the practice in the past 12 months and is summarized below, see:

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability3.html

	Age (years)						
	5	**6**	**7**	**8**	**9**	**10**	**Total**
Boys	432	379	501	410	420	418	2,560
Girls	408	513	412	436	461	500	2,730
Totals	840	892	913	846	881	918	5,290

1.3.1 Unconditional Probability

 **Case Study 2**

If we select a child at random (by simple random sampling), then each child has the same probability (equal chance) of being selected, and the probability is 1/N, where N=the population size. Thus, the probability that any child is selected is 1/5,290 = 0.0002. In most sampling situations we are generally not concerned with sampling a specific individual but instead we concern ourselves with the probability of sampling certain types of individuals. For example, what is the probability of selecting a boy or a child 7 years of age?

The following formula can be used to compute probabilities of selecting individuals with specific attributes or characteristics.

P(characteristic) = # persons with characteristic / N

Try to figure these out before looking at the answers:

1. What is the probability of selecting a boy? **Answer**
2. What is the probability of selecting a 7 year-old? **Answer**
3. What is the probability of selecting a boy who is 10 years of age? **Answer**
4. What is the probability of selecting a child (boy or girl) who is at least 8 years of age? **Answer**

4.3.2 Conditional Probability

 **Case Study 3**

Each of the probabilities computed in the previous section (e.g., P(boy), P(7 years of age)) is an unconditional probability, because the denominator for each is the total population size (N=5,290) reflecting the fact that everyone in the entire population is eligible to be selected. However, sometimes it is of interest to focus on a particular subset of the population (e.g., a sub-population).

For example, suppose we are interested just in the girls and ask the question, what is the probability of selecting a 9 year old from the sub-population of girls?

There is a total of N_G=2,730 girls (here N_G refers to the population of girls), and the probability of selecting a 9 year old from the sub-population of girls is written as follows

P(9 year old | girls) = # persons with characteristic / N
where **| girls** indicates that we are conditioning the question to a specific subgroup, i.e., the subgroup specified to the right of the vertical line.

The conditional probability is computed using the same approach we used to compute unconditional probabilities. In this case:

P(9 year old | girls) = 461/2,730 = 0.169.
This also means that 16.9% of the girls are 9 years of age. Note that this is *not* the same as the probability of selecting a 9-year old girl from the overall population, which is P(girl who is 9 years of age) = 461/5,290 = 0.087.

What is the probability of selecting a boy from among the 6 year olds?

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability3.html

4.3.3 Types of Probabilities

There are three major types of probabilities:
- Theoretical Probability.
- Experimental Probability.
- Axiomatic Probability.

Probabilities can also be classified under two types of probability distribution which are used for different purposes and various types of the data generation process.
1. Normal or Continuous Probability Distribution.
2. Binomial or Discrete Probability Distribution

4.3.4 Probability Model

A probability model is a mathematical representation of a random phenomenon. It is defined by its sample space, events within the sample space, and probabilities associated with each event. The sample space S for a probability model is the set of all possible outcomes.

4.3.5 Reasons for using Probability

Probability provides information about the likelihood that something will happen. Statisticians, for instance, use National Bureau of Statistics data patterns to predict the probability of the state of economy. In management, probability theory is used to understand the relationship between exposures and the risk management, returns and risks, price and demand and so on.

SELF-ASSESSMENT EXERCISE(S)

1.	Explain the word probability
2.	Differentiate between conditional and unconditional probability
3.	What are the 3 types of probability?
4.	What are probability models?
5.	What are the types of probability distributions?

4.4 Summary

In common usage, the word "probability" is used to mean the chance that a particular event (or set of events) will occur expressed on a linear scale from 0 (impossibility) to 1 (certainty), also expressed as a percentage between 0 and 100%. The analysis of events governed by probability is called statistics.

Probabilities and probability theories were considered in cognizant that the number that reflects the chance or likelihood that a particular event will occur.

4.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Alreck, P.L. and Settle, R. (2003). Survey Research Handbook, 3rd edition, McGraw Hill, New York.

Barnes, S. (Ed.) (2005). News of the World, Football Annual 2005–2006, Invincible Press, London, ISBN 0-00-720582-1.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

MODULE 5 STATISTICAL DISTRIBUTIONS

UNIT 1 NORMAL DISTRIBUTION AND STUDENTS (T) DISTRIBUTION

Unit Structure

1.1 Introduction

A statistical distribution is a parameterized mathematical function that gives the probabilities of different outcomes for a random variable. There are discrete and continuous distributions depending on the random value it models.

1.2 Intended Learning Outcomes (ILOs)

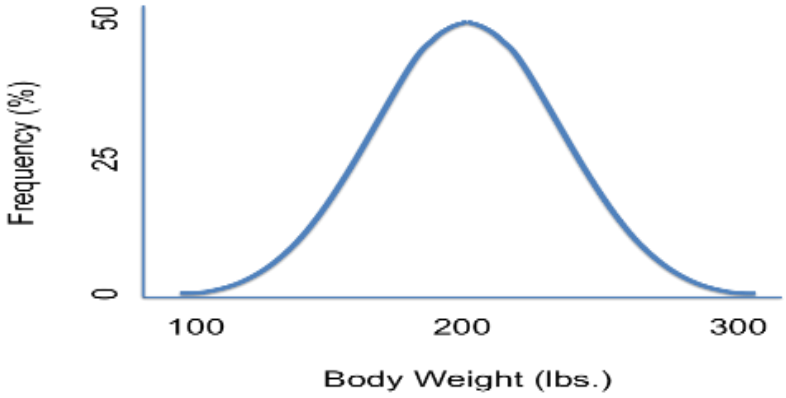
To establish the probabilities of different outcomes for a random variable.

1.3 Main Content

1.3.1 Normal Distribution

1.3.1.1 The Normal Distribution: A Probability Model for a Continuous Outcome

Normal (Gaussian) Distributions



Suppose we were interested in characterizing the variability in body weights among adults in a population. We could measure each subject's weight and then summarize our findings with a graph that displays different body weights on the horizontal axis (the x-axis) and the frequency (% of subjects) of each weight on the vertical axis (the y-axis) as shown in the illustration above.

There are several noteworthy characteristics of this graph. It is bell-shaped with a single peak in the center, and it is symmetrical. If the distribution is perfectly symmetrical with a single peak in the center, then the mean value, the mode, and the median will be all the same. Many variables have similar characteristics, which are characteristic of so-called normal or Gaussian distributions.

Note that the horizontal or x-axis displays the scale of the characteristic being analyzed (in this case weight), while the height of the curve reflects the probability of observing each value. The fact that the curve

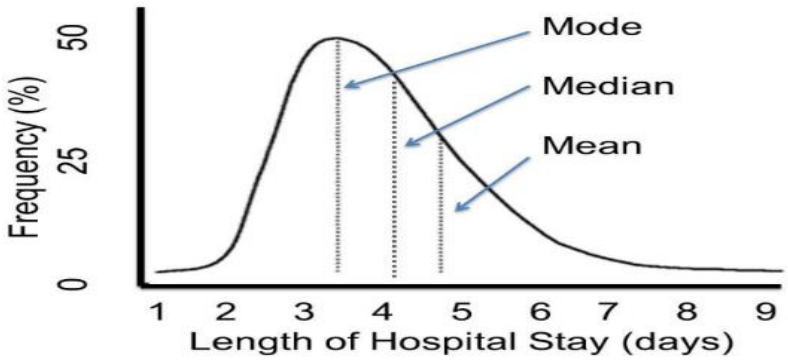
is highest in the middle suggests that the middle values have higher probability or are more likely to occur, and the curve tails off above and below the middle suggesting that values at either extreme are much less likely to occur. There are different probability models for continuous outcomes, and the appropriate model depends on the distribution of the outcome of interest.

The normal probability model applies when the distribution of the continuous outcome conforms reasonably well to a normal or Gaussian distribution, which resembles a bell shaped curve. Also note that normal probability model can be used even if the distribution of the continuous outcome is not perfectly symmetrical; it just has to be reasonably close to a normal or Gaussian distribution.

1.3.2 Skewed Distributions

There are however other distributions which do not follow the symmetrical patterns shown above. For example, if we were to study hospital admissions and the number of days that patients are admitted in the hospital, we would find that the distribution may not be symmetrical, but skewed.

Note that the distribution to the distribution below is not symmetrical, and the mean value is not the same as the mode or the median.



1.3.3 Characteristics of Normal Distributions

Distributions that are normal or Gaussian have the following characteristics:

- Approximately 68% of the values fall between the mean and one standard deviation (in either direction)
- Approximately 95% of the values fall between the mean and two standard deviations (in either direction)
- Approximately 99.9% of the values fall between the mean and three standard deviations (in either direction)

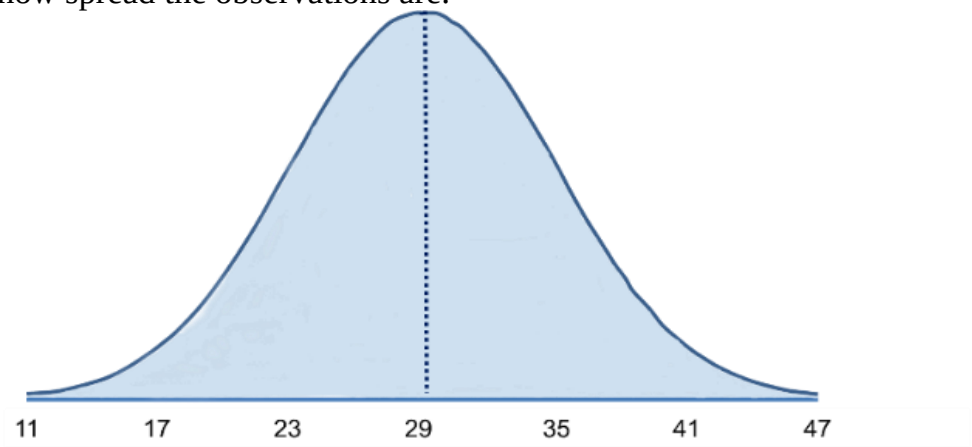
If we have a normally distributed variable and know the population mean, μ and the standard deviation (σ), then we can compute the probability of particular values based on this equation for the normal probability model:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/\left(2\sigma^2\right)}$$

where **μ** is the population mean and **σ** is the population standard deviation. (**π** is a constant = 3.15159, and **e** is a constant = 2.71828.) Normal probabilities can be calculated using calculus or from an Excel spreadsheet. There are also very useful tables that list the probabilities.

Case Study 1 Example of Body Mass Index (BMI) in Males

Consider body mass index (BMI) in a population of 60 year old males in whom BMI is normally distributed and has a mean value = 29 and a standard deviation = 6. The standard deviation gives us a measure of how spread the observations are.

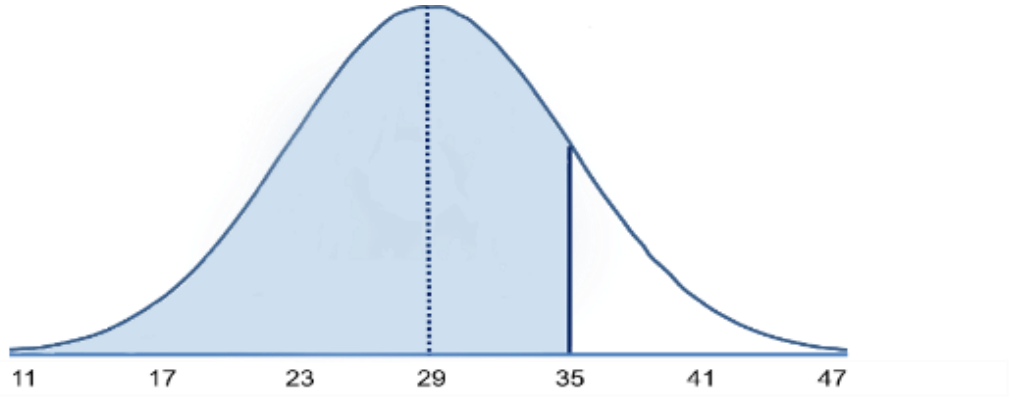


The mean (μ = 29) is in the center of the distribution, and the horizontal axis is scaled in increments of the standard deviation (σ = 6) and the distribution essentially ranges from μ - 3 σ to μ + 3σ. It is possible to have BMI values below 11 or above 47, but extreme values occur very infrequently.

To compute probabilities from normal distributions, we will compute areas under the curve which for any probability distribution is 1. For the normal distribution, we know that the mean is equal to median, so half (50%) of the area under the curve is above the mean and half is below, so P(BMI < 29) = 0.50. Consequently, if we select a man at random from this population and ask of the probability that his BMI is less than 29, his answer will be 0.50 or 50%, since 50% of the area under the curve is below the value BMI = 29.

Note that with the normal distribution the probability of having any exact value is 0 because there is no area at an exact BMI value, so in this case, the probability that his BMI = 29 is 0, but the probability that his BMI is <29 or the probability that his BMI is $\leq$ 29 is 50%.

What is the probability that a 60 year old man has BMI less than 35? The probability is displayed graphically and represented by the area under the curve to the left of the value 35 in the figure below.



Note that BMI = 35 is 1 standard deviation above the mean. For the normal distribution we know that approximately 68% of the area under the curve lies between the mean plus or minus one standard deviation. Therefore, 68% of the area under the curve lies between 23 and 35. We also know that the normal distribution is symmetric about the mean, therefore P (29 < X < 35) = P (23 < X < 29) = 0.34. Consequently, P(X < 35) = 0.5 + 0.34 = 0.84. In other words, 68% of the area is between 23 and 35, so 34% of the area is between 29 and 35, and 50% is below 29. If the total area under the curve is 1, then the area below 35 is therefore, 0.50 + 0.34 = 0.84 or 84%.

Case Study 2

What is the probability that a 60 year old man has BMI less than 41?
[Hint: A BMI of 41 is 2 standard deviations above the mean.] Try to figure this out on your own before looking at the answer.

Solution

It is easy to figure out the probabilities for values that are increments of the standard deviation above or below the mean, but what if the value isn't an exact multiple of the standard deviation? For example, suppose we want to compute the probability that a randomly selected man has a BMI less than 30 (which is the threshold for classifying someone as obese).

Because 30 is neither the mean nor a multiple of standard deviations above or below the mean, we cannot simply use the probabilities known to be associated with 1, 2, or 3 standard deviations from the mean. In a sense, we need to know how far a given value is from the mean and the probability of having values less than this. And, of course, we would want to have a way of figuring this out not only for BMI values in a population of males with a mean of 29 and a standard deviation of 6, but for any normally distributed variable. So, what we need is a standardized way of evaluating any normally distributed data so that we can compute the probability of observing the results obtained from samples that we take. We can do all of this by using a "standard normal distribution.

1.3.4 Z Scores are Standardized Scores

 **Case Study 3**

Assuming that the body mass index (BMI) in a population of 60 year old man's BMI is normally distributed with mean value = 29 and a standard deviation = 6.

What is the probability that a randomly selected man from this population would have a BMI less than 30?

While a value of 30 doesn't fall on one of the increments of standard deviation, we can calculate how many standard deviations it is away from the mean.

It is 30-29=1 BMI unit above the means. The standard deviation is 6, so 1 BMI unit above the mean is1/6 = 0.156667 standard deviations above the mean. This shows the distance a given observation is from the mean for any normal distribution, regardless of its mean or standard deviation. The next step is to find a way of solving the probabilities associated various Z-scores. This can be done by using the standard normal distribution.

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability8.html

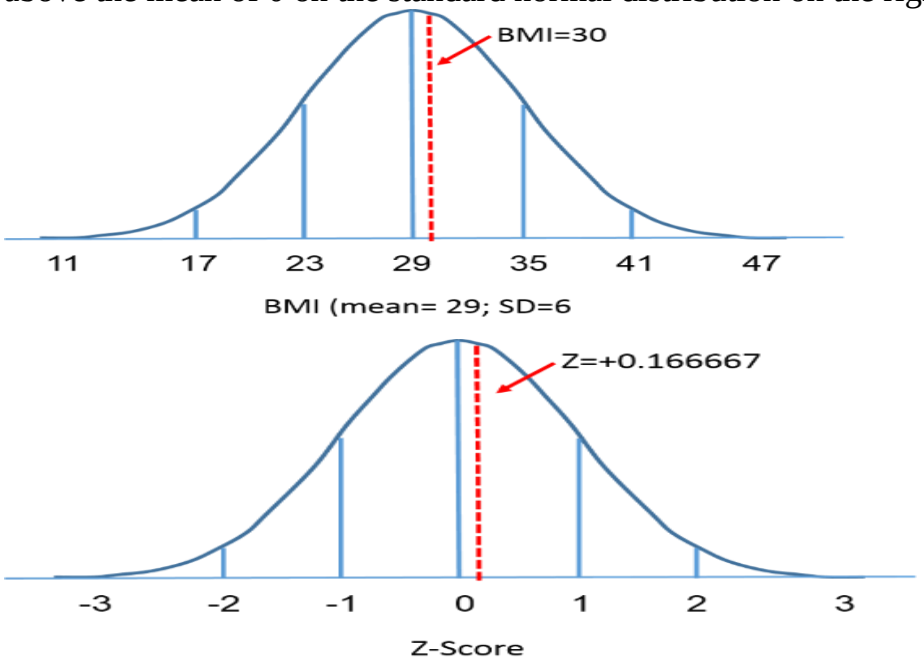
1.3.5 The Standard Normal Distribution

The standard normal distribution is a normal distribution with a mean of zero and standard deviation of 1. The standard normal distribution is centred at zero and the degree to which a given measurement deviates from the mean is given by the standard deviation.

1.3.6 Basics for the standard normal distribution;

68% of the observations lie within 1 standard deviation of the mean; 95% lie within two standard deviations of the mean; and 99.9% lie within 3 standard deviations of the mean.

To this point, we have been using "X" to denote the variable of interest (e.g., X=BMI, X=height, X=weight). However, when using a standard normal distribution, "Z" is used to denote a variable in the context of a standard normal distribution. After standardization, the BMI=30 discussed on the previous page is shown below lying 0.15667 units above the mean of 0 on the standard normal distribution on the right.



Since the area under the standard curve = 1, is easy to obtain the probabilities of specific observation. For any given Z-score we can compute the area under the curve to the left of that Z-score.it is pertinent to note that a "Z" score of 0.0 tilts a probability of 0.50 or 50%, and a "Z" score of 1, meaning one standard deviation above the mean, lists a probability of 0.8413 or 84%. That is because one standard deviation above and below the mean encompasses about 68% of the area, so one standard deviation above the mean represents half of that of 34%. So, the 50% below the mean plus the 34% above the mean gives us 84%.

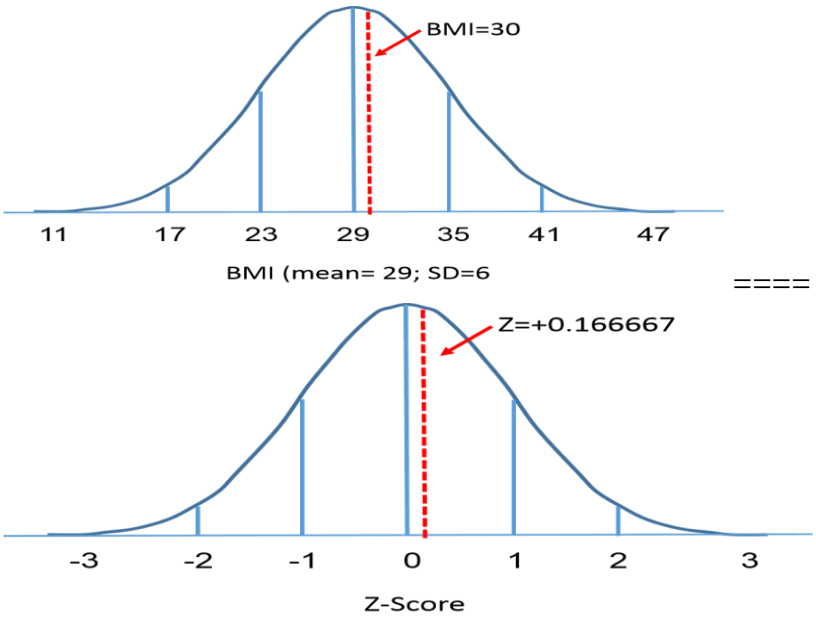
1.3.6 Probabilities of the Standard Normal Distribution, Z

The statistical normal table is organized to provide the area under the curve to the left of or less of a specified value or "Z value". In this case, because the mean is zero and the standard deviation is 1, the Z value is

the number of standard deviation units away from the mean, and the area is the probability of observing a value less than that particular Z value. Note also that the normal table shows probabilities to two decimal places of Z. The units place and the first decimal place are shown in the left hand column, and the second decimal place is displayed across the top row.

But let's get back to the question about the probability that the BMI is less than 30, that is, P(X<30). We can answer this question using the standard normal distribution. The figures below show the distributions of BMI for men aged 60 and the standard normal distribution side-by-side.

1.3.7 Distribution of BMI and Standard Normal Distribution



The area under each curve is one but the scaling of the X axis is different. Note, however, that the areas to the left of the dashed line are the same. The BMI distribution ranges from 11 to 47, while the standardized normal distribution, Z, ranges from -3 to 3. We want to compute P(X < 30). To do this we can determine the Z value that corresponds to X = 30 and then use the standard normal distribution table above to find the probability or area under the curve. The following formula converts an X value into a **Z score**, also called a **standardized score**:

$Z = \frac{X-\mu}{\sigma}$; where μ is the mean and σ is the standard deviation of the variable X.

Case Study 4

In order to compute P(X < 30) we convert the X=30 to its corresponding Z score (this is called **standardizing**):

$$Z = \frac{30-29}{6} = \frac{1}{6} = 0.17$$

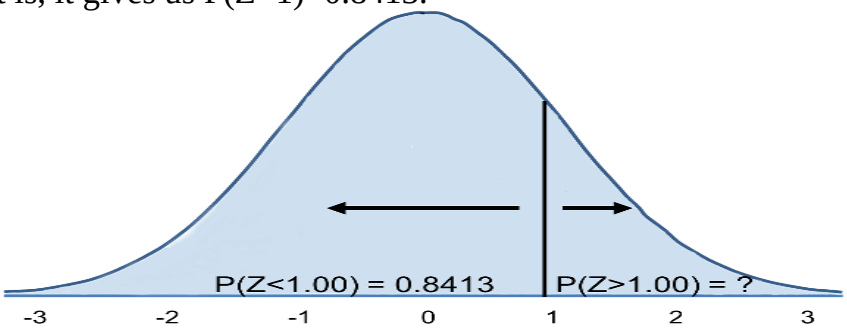
Thus, P(X < 30) = P(Z < 0.17). We can then look up the corresponding probability for this Z score from the standard normal distribution table, which shows that P(X < 30) = P(Z < 0.17) = 0.5675. Thus, the probability that a male aged 60 has BMI less than 30 is 56.75%.

Case Study 5

Using the same distribution for BMI, what is the probability that a male aged 60 has BMI **exceeding** 35? In other words, what is P(X > 35)? Again we standardize:

$$Z = \frac{35-29}{6} = \frac{6}{6} = 1$$

We now go to the standard normal distribution table to look up P(Z>1) and for Z=1.00 we find that P(Z<1.00) = 0.8413. Note, however, that the table always gives the probability that Z is ***less*** than the specified value, that is, it gives us P(Z<1)=0.8413.



Therefore, P(Z>1)=1-0.8413=0.1587. Interpretation: Almost 15% of men aged 60 have BMI over 35.

1.3.8 Computing Percentiles

The standard normal distribution can also be useful for computing **percentiles**. For example, the median is the 50th percentile, the first quartile is the 25th percentile, and the third quartile is the 75th percentile. In some instances it may be of interest to compute other percentiles, for example the 5th or 95th. The formula below is used to compute percentiles of a normal distribution.

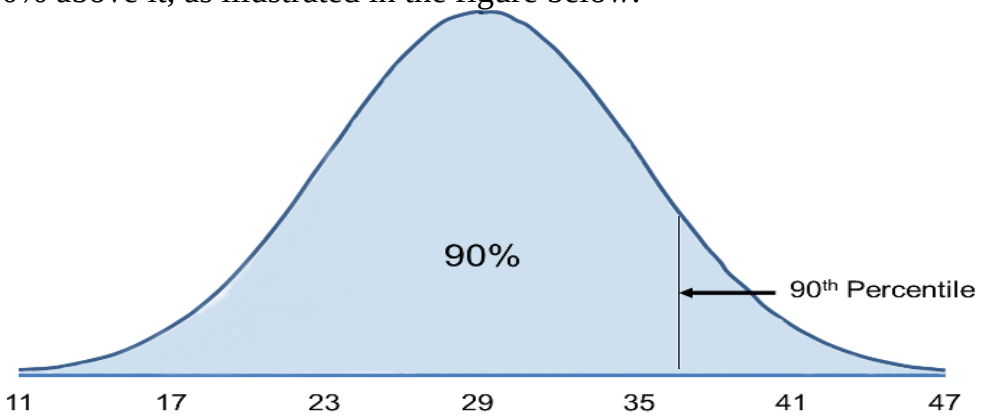
$$X = \mu + Z\sigma$$

where μ is the mean and σ is the standard deviation of the variable X, and Z is the value from the standard normal distribution for the desired percentile.

Case Study 6

- The mean BMI for men aged 60 is 29 with a standard deviation of 6.
- The mean BMI for women aged 60 the mean is 28 with a standard deviation of 7.

What is the 90th percentile of BMI for men?
The 90th percentile is the BMI that holds 90% of the BMIs below it and 10% above it, as illustrated in the figure below.



To compute the 90th percentile, given the formula $X=\mu + Z\sigma$, and standard normal distribution table, unlike the previous example on BMI, computation will be done in the opposite direction. Previously we started with a particular "X" and used the table to find the probability. However, in this case we want to start with a 90% probability and find the value of "X" that represents it.

So we begin by going into the interior of the standard normal distribution table to find the area under the curve closest to 0.90, and from this we can determine the corresponding Z score. Once we have this we can use the equation $X=\mu + Z\sigma$, because we already know that the mean and standard deviation are 29 and 6, respectively.

When we go to the table, we find that the value 0.90 is not there exactly, however, the values 0.8997 and 0.9015 are there and correspond to Z values of 1.28 and 1.29, respectively (i.e., 89.97% of the area under the standard normal curve is below 1.28). The exact Z value holding 90% of the values below it, is 1.282 which was determined from a table of standard normal probabilities with more precision.

Using Z=1.282 the 90th percentile of BMI for men is: X = 29 + 1.282(6) = 36.69.

Interpretation: Ninety percent of the BMIs in men aged 60 are below 36.69. Ten percent of the BMIs in men aged 60 are above 36.69.

 **Case Study 7**

What is the 90th percentile of BMI among women aged 60? Recall that the mean BMI for women aged 60 the mean is 28 with a standard deviation of 7.

Solution

The table below shows Z values for commonly used percentiles.

Percentile	**Z**
1st	-2.326
2.5th	-1.960
5th	-1.645
10th	-1.282
25th	-0.675
50th	0
75th	0.675
90th	1.282
95th	1.645
97.5th	1.960
99th	2.326

Percentiles of height and weight are used by pediatricians in order to evaluate development relative to children of the same sex and age.

 **Case Study 8**

If a child's weight for age is extremely low it might be an indication of malnutrition.

1. For infant girls, the mean body length at 10 months is 72 centimeters with a standard deviation of 3 centimeters. Suppose a girl of 10 months has a measured length of 67 centimeters. How does her length compare to other girls of 10 months?
2. A complete blood count (CBC) is a commonly performed test. One component of the CBC is the white blood cell (WBC) count, which may be indicative of infection if the count is high. WBC counts are approximately normally distributed in healthy people

with a mean of 7550 WBC per mm^3 (i.e., per microliter) and a standard deviation of 1085. What proportion of subjects have WBC counts exceeding 9000?

3. Using the mean and standard deviation in the previous question, what proportion of patients have WBC counts between 5000 and 7000?

The table of areas under the normal curve shows the area between the mean and a given number of standard deviation. Recall that if X is a continuous random variable, X is said to be normally distributed with parameter μ *and* σ^2 $(X \sim N(\mu, \sigma^2)$ with pdf given as;

$$f(x) = \frac{1}{\sqrt[\sigma]{2A}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Where μ = population mean, σ = standard deviation, λ = pie (3.152), ℓ = the base of natural logarithm.

The total area bounded by the normal curve and X is 1. Hence, the area under the curve between two coordinate X=A and X=B is denoted by Pr(a<X<b) .When the variable X is to be expressed standardized form as; $Z = \frac{X - \mu}{\sigma}$.

The equation above is replaced by $f(x) = \frac{1}{2\pi} e^{-\frac{1}{2}Z^2}$.

In such a case, we say that Z is normally distributed with μ as zero and σ as 1 (0,1), that is, Z~N(0,1), where Z is the standardize form of X. Since the pdf of the standard normal distribution is symmetric with respect to point Z=0 (zero). It follows that $p(Z \le Z) = p(Z \ge -Z)$.

For any real number $Z(-\infty < Z < \infty)$ .The mean to the right of Z is equal to zero (o) and the area to the left of Z=0 is 0.5.

1.3.9 Evaluation of Probabilities for a Normal Distribution

 Case Study 9**:**

i. $p(-t \le Z \le S)$

$p(-t \le Z \le S) = A(t) + A(S)$ is shaded area.

ii. $p(Z \ge S)$

iii. $p(Z \le S)$

$p(Z \le S) = 0.5 + A(S)$

If Z follows normal (0,1) find

i. P(Z>-0.5)

Solution

i. P(Z>-0.5)

P(Z>-0.5)=0.5+0.1915=0.6915

 **Case Study 10;**

Suppose that the birth weight of a baby is normally distributed with μ =3500g, σ =500g . What is the probability that a baby born has weight less that 3100.

Solution

The probability that a baby born has weight less that 3100 P(X<3100)

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

$$P\left[\frac{X - 3500}{500} < \frac{3100 - 3500}{500}\right]$$

$$P\left[Z < \frac{-400}{500}\right]$$

$$P[Z < -0.8]$$

$$p(Z < -0.8) = 0.5 - 0.2881 = 0.2119$$

1.3.10 Students (t) Distribution

The T distribution, also known as the Student's t-distribution, is a type of probability distribution that is similar to the normal distribution with its bell shape but has heavier tails. T distributions have a greater chance for extreme values than normal distributions, hence the fatter tails.

The Student's t-distribution) is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

The t - distribution can be regarded as distribution of function of random variables.

Level of significance is the maximum probability with which we will be willing to risk a type 1 error (error of rejecting a hypothesis that should be accepted).

Degree of freedom is the number of observations (or values) that are independent of each other i.e that can't be deducted from each other.

1.3.11 Fitting a Normal Curve to a Data

Case Study 11

The table below is a frequency distribution of height recorded to the nearest inch of 100 male students at XYZ University

Height	No. of student	X	$f\bar{X}$	$(\bar{X}-\bar{X})$	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
60-62	5	61	305	-6.45	41.603	208.02
63-65	18	64	1152	-3.45	11.903	215.25
66-68	42	67	2815	-0.45	0.203	8.53
69-71	27	70	1890	2.55	6.503	175.58
72-74	8	73	584	5.55	30.803	246.44
	100		6745	-2.25		852.82

a. Fit a normal curve to the data

b. Determine the goodness of fit of the data

Recall $\Rightarrow Z = \dfrac{X-\bar{X}}{S}$

Where $\bar{X}$ is the mean

& s is the S deviation

$$S = \sqrt{\frac{\sum f(X-\bar{X})^2}{\sum f}}$$

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

$$\bar{X} = \frac{\sum fx}{\sum f} = \frac{67.45}{100} = 67.45$$

$$S = \sqrt{\frac{852.82}{100}} = \sqrt{8.5282}$$

S=2.92

Height	C.B(X)	Z for C.B	Area under normal curve 0 to Z	Area for each class	Expected frequencies	Observed frequencies
60-62	59.5-62.5	-2.72	0.4967	0.0411	4.11	5
63-65	62.5-65.5	-1.70	0.4554	0.2068	20.68	18

66-68	65.5-68.5	-0.67	0.2486	0.3892	38.92	42
69-71	68.5-71.5	0.36	0.1506	0.2771	27.71	27
72-74	71.5-74.5	1.39	0.4177	0.0743	7.43	8
		2.41	0.4920			

Hint;

a. Expected frequencies = Area for each class x 100.

b. Use Chi-Square goodness of test to determine the goodness of fit of the data; $X^2 = \dfrac{\sum_i^n (O_i - e_i)^2}{e_i}$

c. Hypothesis;
Null Hypothesis H_o – the fit is good for the data.
Alternative Hypothesis H_1 – the fit is not good for the data.
d. If Chi-Square calculated is lesser than chi-square tabulated, we accept the H_o.
If chi-square calculated is greater than chi-square tabulated, we accept the H_1

$$X^2 = \frac{(5-4.13)^2}{4.13} + \frac{(18-20.68)^2}{20.68} + \frac{(42-38.92)^2}{38.92} + \frac{(27-27.71)^2}{27.71} + \frac{(8-7.43)^2}{7.43}$$

=0.183+0.347+0.244+0.018+0.044 =0.836; $X^2_{cal} = 0.836$

Test using 95%=0.95 and 5%=0.05

Recall; V=K-1-M, where V=degree of freedom, K-number of observations, M-number of parameter used to estimate K=5, M=2 (mean and standard); then V=5-1-2=2 $X^2_{tab} = X^2_{2,0.95}$ ($X^2_{V,level\ of\ significance}$)= 5.99 at 95% level of significance. Since $X^2_{cal} < X^2_{tab}$ we do not reject the null hypothesis (H_o) – meaning the fit is good for the data.

Test again using 0.05 as level of significance; $X^2_{2,0.05} = 0.103$

At 5% level of significance, the $X^2_{cal} > X^2_{tab}$ therefore we accept the alternative hypothesis (H_1) and reject the null, meaning the fit is not good for the data.

1.3.12 Difference between z test and t test

Z-tests are statistical calculations that can be used to compare population means to a sample's. **T-tests** are calculations used to test a hypothesis, but they are most useful when we need to determine

if there is a statistically significant difference between two independent sample groups.

1.3.13 Difference between the t- distribution and the normal distribution

The normal distribution is used when the population distribution of data is assumed normal. A sample of the population is used to estimate the mean and standard deviation. The **t** statistic is an estimate of the standard error of the mean of the population or how well known is the mean based on the sample size.

1.3.14 Application of t-distribution

The **t-distribution** is **used** as an alternative to the normal distribution when sample sizes are small in order to estimate confidence or determine critical values that an observation is a given distance from the mean. As the sample size increases, so do degrees of freedom. When degrees of freedom are infinite, the **t-distribution** is identical to the normal distribution. As sample size increases, the **s**ampl**e** more closely approximates the population.

SELF-ASSESSMENT EXERCISE(S)

1. What do you understand by the following;
   a. normal distribution
   b. students (t) distribution
2. If Z follows normal (0,1) find
   a. P(Z>0.92)
   b. P(Z<-0.76)
   c. P(-0.64<Z<0.43)
3. Suppose that the birth weight of a baby is normally distributed with μ =3500g, σ =500g. What is the probability that a baby born has weight less than 3100?
4. Why does T distribution depend on sample size?
   What is the difference between the T distribution and the normal distribution?

1.4 Summary

Apart from unit 1 establishing the probabilities of different outcomes for a random variable in distributions, it also ascertains the goodness of the test.

This unit have been able to establish the probabilities of different outcomes for a random variable.

1.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Alreck, P.L. and Settle, R. (2003). Survey Research Handbook, 3rd edition, McGraw Hill, New York.

Barnes, S. (Ed.) (2005). News of the World, Football Annual 2005–2006, Invincible Press, London, ISBN 0-00-720582-1.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

UNIT 2 BINOMIAL DISTRIBUTION

Unit Structure

2.1 Introduction

The binomial distribution model allows us to compute the probability of observing a specified number of "successes" when the process is repeated a specific number of times, for example, in a set of patients and the outcome for a given patient is either a success or a failure. We must first introduce some notation which is necessary for the binomial distribution model.

 First, we let "n" denote the number of observations or the number of times the process is repeated, and "x" denotes the number of "successes" or events of interest occurring during "n" observations. The probability of "success" or occurrence of the outcome of interest is indicated by "p".

The binomial equation also uses factorials. In mathematics, the factorial of a non-negative integer k is denoted by k!, which is the product of all positive integers less than or equal to k.
For example,

- 4! = 4 x 3 x 2 x 1 = 24,
- 2! = 2 x 1 = 2,
- 1!=1.
- There is one special case, 0! = 1.

With this notation in mind, the binomial distribution model is defined as:
The Binomial Distribution Model

$$P(X \text{ "successes"}) = \frac{n!}{x!\,(n-x)!} p^{x} (1-p)^{(n-x)}$$

2.2 Intended Learning Outcomes (ILOs)

To compute the probability of a range of outcomes

2.3 Main Content

2.3.1 Definition of Binomial distribution

Binomial distribution summarizes the number of trials, or observations when each trial has the same probability of attaining one particular value. The **binomial distribution** determines the probability of observing a specified number of successful outcomes in a specified number of trials.

2.3.2 The four requirements are:

- each observation falls into one of two categories called a success or failure.
- there is a fixed number of observations.
- the observations are all independent.
- the **probability** of success (p) for each observation is the same - equally likely.

2.3.3 Properties of a binomial experiment

A **binomial** experiment has four **properties**: 1) it consists of a sequence of n identical trials; 2) two outcomes, success or failure, are possible on each trial; 3) the **probability** of success on any trial, denoted p, does not.

2.3.4 Difference between binomial distribution and Poisson distribution

Binomial distribution describes the distribution of binary data from a finite sample. Poisson distribution describes the distribution of binary data from an infinite sample. Thus it gives the probability of getting r events in a population.

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/bs704_probability7.html
A multinomial probability model might be appropriate, but here we focus on the situation in which the outcome is dichotomous.

Case Study 1
Adults with allergies might report relief with medication or not, children with a bacterial infection might respond to antibiotic therapy or not, adults who suffer a myocardial infarction might survive the heart attack or not, a medical device such as a coronary stent might be successfully implanted or not. These are just a few examples of applications or processes in which the outcome of interest has two possible values (i.e., it is dichotomous). The two outcomes are often labelled "success" and "failure" with success indicating the presence of the outcome of interest. Note, however, that for many medical and public health questions the outcome or event of interest is the occurrence of disease, which is obviously not really a success. Nevertheless, this terminology is typically used when discussing the binomial distribution model. As a result, whenever using the binomial distribution, we must clearly specify which outcome is the "success" and which is the "failure".

2.3.5 Use of the binomial distribution requires three assumptions:

1. Each replication of the process results in one of two possible outcomes (success or failure),
2. The probability of success is the same for each replication, and
3. The replications are independent, meaning here that a success in one patient does not influence the probability of success in another.

Examples of Use of the Binomial Model

 **Case Study 2 Relief of Allergies**

Suppose that 80% of adults with allergies report symptomatic relief with a specific medication. If the medication is given to 10 new patients with allergies, what is the probability that it is effective in exactly seven patients?

First, do we satisfy the three assumptions of the binomial distribution model?

- ✓ The outcome is relief from symptoms (yes or no), and here we will call a reported relief from symptoms a 'success.'
- ✓ The probability of success for each person is 0.8.
- ✓ The final assumption is that the replications are independent, and it is reasonable to assume that this is true.

We know that:

- number observation is n=10
- number successes or events of interest is x=7
- P = 0.80

The probability of 7 successes is:

$$P(7 \text{ successes}) = \frac{10!}{7!(10-7)!} 0.80^{7} (1-0.80)^{(10-7)}$$

This is equivalent to:

$$P(7 \text{ successes}) = \frac{10(9)(8)(7)(6)(5)(4)(3)(2)(1)}{[7(6)(5)(4)(3)(2)(1)]\ [(3)(2)(1]} (0.80)^{7} (1-0.80)^{10-7}$$

But many of the terms in the numerator and denominator cancel each other out,

$$P(7 \text{ successes}) = \frac{10(9)(8)}{3(2)(1)} (0.2097)(0.008) = 120(0.0297)(0.008) = 0.2013$$

Interpretation: There is a 20.13% probability that exactly 7 of 10 patients will report relief from symptoms when the probability that any one reports relief is 80%.

 :

Binomial probabilities like this can also be computed in an Excel spreadsheet using the BINOMDIST function. Place the cursor into an empty cell and enter the following formula:

BINOMDIST(x, n, p, FALSE)

where x= # of 'successes', n = # of replications or observations, and p = probability of success on a single observation.What is the probability

that none report relief? We can again use the binomial distribution model with n=10, x=0 and p=0.80.

$$P(\mathbf{0\ successes}) = \frac{10!}{0!(10-0)!}0.80^{0}(1-0.80)^{10-0}$$

This is equivalent to

$$P(\mathbf{0\ successes}) = \frac{10!}{(1)(10)!}0.80^{0}(0.20)^{10}$$

Which simplifies to

$$P(\mathbf{0\ successes}) = (1)\ (1)\ (0.0000001024) = 0.0000001024$$

Interpretation: There is practically no chance that none of the 10 will report relief from symptoms when the probability of reporting relief for any individual patient is 80%.

What is the most likely number of patients who will report relief out of 10? If 80% report relief and we consider 10 patients, we would expect that 8 report relief. What is the probability that exactly 8 of 10 report relief? We can use the same method that was used above to demonstrate that there is a 30.30% probability that exactly 8 of 10 patients will report relief from symptoms when the probability that any one reports relief is 80%. The probability that exactly 8 report relief will be the highest probability of all possible outcomes (0 through 10).

Case Study 3: The Probability of Dying after a Heart Attack

The likelihood that a patient with a heart attack dies of the attack is 0.04 (i.e., 4 of 100 die of the attack). Suppose we have 5 patients who suffer a heart attack, what is the probability that all will survive? For this example, we will call a success a fatal attack (p = 0.04). We have n=5 patients and want to know the probability that all survive or, in other words, that none are fatal (0 successes).

We again need to assess the assumptions. Each attack is fatal or non-fatal, the probability of a fatal attack is 4% for all patients and the outcome of individual patients are independent. It should be noted that the assumption that the probability of success applies to all patients must be evaluated carefully. The probability that a patient dies from a heart attack depends on many factors including age, the severity of the attack, and other comorbid conditions. To apply the 4% probability we must be convinced that all patients are at the same risk of a fatal attack. The assumption of independence of events must also be evaluated carefully. As long as the patients are unrelated, the assumption is usually appropriate. Prognosis of disease could be related or correlated in members of the same family or in individuals who are cohabiting. In this

example, suppose that the 5 patients being analyzed are unrelated, of similar age and free of comorbid conditions.

$$P(\textbf{0 successes}) = \frac{5!}{0!(5-0)!}0.04^{0}(1-0.04)^{5-0}$$

$$P\left(\textbf{0 successes}\right) = \frac{5!}{5!}(1)\ (0.96)^{5} = \ (1)\ (1)\ (0.8154) = 0.8154$$

There is an 81.54% probability that all patients will survive the attack when the probability that any one dies is 4%. In this example, the possible outcomes are 0, 1, 2, 3, 4 or 5 successes (fatalities). Because the probability of fatality is so low, the most likely response is 0 (all patients survive). The binomial formula generates the probability of observing exactly x successes out of n.

16.6 Computing the Probability of a Range of Outcomes

If we want to compute the probability of a range of outcomes we need to apply the formula more than once. Suppose in the heart attack example we wanted to compute the probability *that* no more than 1 person dies of the heart attack. In other words, 0 or 1, but not more than 1. Specifically we want P(no more than 1 success) = P(0 or 1 successes) = P(0 successes) + P(1 success). To solve this probability we apply the binomial formula twice.

We already computed P(0 successes), we now compute P(1 success):

$$P(\textbf{1 success}) = \frac{5!}{1!(5-1)!}0.04^{1}(1-0.04)^{5-1}$$

$$P(\textbf{1 success}) = \frac{5!}{(1)(4)!}(0.04)\ (0.96)^{4} = (5)(0.04)\ (0.8493) = 0.1697$$

P(no more than 1 'success') = P(0 or 1 successes) = P(0 successes) + P(1 success)

= 0.8154 + 0.1697 = 0.9851.

The probability that no more than 1 of 5 (or equivalently that at most 1 of 5) die from the attack is 98.51%.

What is the probability that 2 or more of 5 die from the attack? Here we want to compute P(2 or more successes). The possible outcomes are 0, 1, 2, 3, 4 or 5, and the sum of the probabilities of each of these outcomes is 1 (i.e., we are certain to observe either 0, 1, 2, 3, 4 or 5 successes). We just computed P(0 or 1 successes) = 0.9851, so P(2, 3, 4 or 5 successes) = 1 - P(0 or 1 successes) = 0.0149. There is a 1.49% probability that 2 or more of 5 will die from the attack.

2.3.3 Mean and Standard Deviation of a Binomial Population

 **Case Study 4**

Mean number of successes: $\mu = np$

Standard Deviation: $\sigma = \sqrt{n\,(p)\,(1-p)}$

For the previous example on the probability of relief from allergies with n-10 trials and p=0.80 probability of success on each trial:

$\mu = np = (10)\,(0.80) = 8$

$\sigma = \sqrt{n\,(p)\,(1-p)} = \sqrt{10\,(0.8)\,(0.2)} = 1.3$

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/bs704_probability7.htm

SELF-ASSESSMENT EXERCISE(S)

1. What is Binomial distribution?
2. What are the four requirements of Binomial distribution?
3. What are the properties of a binomial experiment?
4. What are the differences between binomial and Poisson distributions?
5. When do we apply binomial distribution?

 **2.4 Summary**

In binomial process each replication of the process results in one of two possible outcomes, that is, success or failure and the probability of success is the same for each replication. The replications of binomial distributions are independent.

Unit 2 computed the probability of a range of outcomes

2.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Alreck, P.L. and Settle, R. (2003). Survey Research Handbook, 3rd edition, McGraw Hill, New York.

Barnes, S. (Ed.) (2005). News of the World, Football Annual 2005–2006, Invincible Press, London, ISBN 0-00-720582-1.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/bs704_probability7.htm

UNIT 3 POISSON, GEOMETRIC AND HYPER-GEOMETRIC DISTRIBUTIONS

Unit Structure

3.1 Introduction

In probability theory and statistics, the Poisson distribution was named after French mathematician Simeon Denis Poisson. Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space, if these events occur with a known constant mean rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

3.2 Intended Learning Outcomes (ILOs)

To establish that the discrete probability distribution of the number of events occurring in a given time period, given the average number of times the event occurs over that time period.

3.3 Main Content

3.3.1 Definition of Poisson distribution

The Poisson distribution is a discrete distribution that measures the probability of a given number of events happening in a specified time period. It is also the number of events occurring in a given time period, given the average number of times the event occurs over that time period.

In finance, the Poisson distribution could be used to model the arrival of new buy or sell orders entered into the market or the expected arrival of orders at specified trading venues or dark pools. In these cases, the Poisson distribution is used to provide expectations surrounding confidence bounds around the expected order arrival rate. Poisson distributions are very useful for smart order routers and algorithmic trading.

The Poisson distribution describes the probability to find exactly *x* events in a given length of time if the events occur independently at a constant rate. In addition, the Poisson distribution can be obtained as an approximation of a binomial distribution when the number of trials *n* of the latter distribution is large, success probability *p* is small, and *np* is a finite number.

3.3.2 Probability of events for a Poisson distribution

A event can occurs 0,1,2,…times in an interval. The average number of events in an interval is denoted, λ. Λ is the events rate, also called the rate parameter. The probability of observing k events in an interval (the PMF of the Poisson distribution) is given by the equation.

$$P(k,\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}\text{; For x= 0, 1, 2, …} \quad 3.1$$

where λ is a positive number or the average number of events per interval.
e is the number 2.71728 (the base of the natural logarithms)
K takes value 0,1,2,….

$$k! - k \times (k-1) \times (k-2) \ldots \times 2 \times 1 \; is\; the\; factorial\; of\; k \quad 3.2$$

Both the mean and variance of the Poisson distribution are equal to λ.

3.3.3 Characteristics of a Poisson distribution

The experiment consists of counting the number of events that will occur during a specific interval of time or in a specific distance, area, or volume.

The difference between Poisson and binomial distribution
The Binomial and Poisson distributions are similar, but they are different. The difference between the two is that while both measure the number of certain random events (or "successes") within a certain frame, the Binomial is based on discrete events, while the Poisson is based on continuous events.

3.3.3.1 Application of Poisson

The Poisson distribution is used to describe the distribution of rare events in a large population. For example, at any particular time, there is a certain probability that a particular cell within a large population of cells will acquire a mutation. Mutation acquisition is a rare event.

3.3.3.2 How do you know if a distribution is Poison?

If a mean or average probability of an event happening per unit time/per page/per mile cycled etc., is given, and you are asked to calculate a probability of n events happening in a given time/number of pages/number of miles cycled, then the Poisson Distribution is used.

Case Study 1

Given that 5% of a population are left handed. Use the Poisson distribution to estimate the probability that a random sample of 100 people contains 2 or more left-handed people
X – number of left-handed people in a sample of 100

$$X \sim Bin(100, 0{,}05)$$

Poisson approximation $X \sim p(\lambda)\ with\ \lambda = 100 \times 0.05 = 5$

$$\text{we want } P(X \geq 2) = 1 - P(X < 2)$$

$$1 - P(X = 0) + (X = 1)$$

$$1 - \left(e^{-5}\frac{5^0}{0!} + e^{-5}\frac{5^1}{1!}\right) = 1 - 0.040428 = 0.959572$$

 **Case Study 2**

What is the probability of selecting four accidents randomly within a week?

$$P(X = 4) = \frac{e^{-\lambda} \times 2^4}{4!} = \frac{0.1353 \times 17}{4 \times 3 \times 2}$$

 **Case Study 3**

X = exactly 50 unique visitors/1hour
E = 2.71728
X= 2.9unique visitor/4 minutes.

First step is to get the units of time to match.
x = exactly 50 unique visitors/1hour
X = 43.5 unique visitors/1hour
(Multiply both numerator and denominator by 15) substitute the Poisson probability equation;

$$P(X) = \frac{e^{-43.5} \times 43.5^{50}}{50!} = 0.035$$

3.3.4 Geometric Distribution

The geometric distribution represents the number of failures before you get a success in a series of Bernoulli trials. This discrete probability distribution is represented by the probability density function:

$f(x) = (1 - p)^{x-1}p$

It is a special case of the negative binomial distribution. It deals with the number of trials required for a single success. Thus, the geometric distribution is a negative binomial **distribution** where the number of successes (r) is equal to 1.

3.3.4.2 Criteria for a distribution to be geometric

(1) The chance experiment must only have two outcomes (success/failure) per trial,
(2) the trials must be independent,
(3) there must be a fixed probability of success for each trial, and

(4) the variable of interest is the number of trials needed to obtain a "success

3.3.5 Difference between binomial and geometric distribution

The basic difference between binomial and geometric distribution is that Binomial: has a FIXED number of trials before the experiment begins and X counts the number of successes obtained in that fixed number. Geometric: has a fixed number of successes (ONE...the FIRST) and counts the number of trials needed to obtain that first success.

3.3.6 Geometric probability formula

If X has a geometric with probability P of success and (1-p) of failure on each observation, the possible value of X are 1, 2, 3, …
If n is any one of these values, the probability that the first success occurs on the nth trial is

$$P(X = n) = P(1 - P)^{n-1} \quad 3.4$$

The probability that it takes more than n trials is

$$P(X > n) = (1 - P)^{n} \quad 3.5$$

Case Study 4

A certain basket player has a 65% chance of making a free throw. Assume all free throws are independent. What is the probability that he makes the first free throws on the 3rd try?
Solution

a. $P(P = 3) = (35^2)65 = 0.79325\,; P = 65\ and\ q = 35$

Case Study 5

How many tossed of a pair of fair die are necessary to be 99% certain that a double size will appear? Find c ∋ $P(x \leq c) \geq 0.99\ where\ x = 1, 2, 3, \ldots$

$$f(x) = (\frac{1}{36} \times \frac{35}{36})^{x-1}{}'$$

$= \sum_{x-1}^{c}(\frac{1}{36})(^{35}/_{36})^{x-1} \geq 0.99$

$=1\text{-}\sum_{x-1}^{c}(\frac{1}{36})(^{35}/_{36})^{x-1} \geq 0.99$

Case Study 6

A fast food chain puts a winning game piece on every fifth package of French fries. Find the probability that you will win a price

a. with your third purchase of French fries
b. with your third or fourth purchase of French fries.

Solutions;

a. $x = 3$

$P(3) = (0.2)(0.8)^{3-1} = 0.128$

b. $x = 3,4;\ P(3or4) = P(3) - P(4) = 0.128 + 0.102 = 0,230$

3.3.7 Hypergeometric Distribution

Definition of Hypergeometric Distribution

In probability theory and statistics, the hypergeometric distribution is a discrete probability distribution that describes the probability of k successes (random draws for which the object drawn has a specified feature) in n draws, *without* replacement, from a finite population of size N that contains exactly K objects with that feature, wherein each draw is either a success or a failure.

In contrast, the binomial distribution describes the probability of k successes in n draws *with* replacement.

FORMULA

A random variable x follows the hypergeometric distribution if its probability mass function (pmf) is given by

$$P(x) = \frac{C_k^K \times C_{n-k}^{N-K}}{C_n^N} \qquad 3.6$$

where

- K is the population size,
- K is the number of success samples in the population,
- n is the number of draws (i.e. quantity drawn in each trial),
- N is the number of observed successes,

3.3.8 What is the hyper-geometric distribution used for?

 The **hyper-geometric distribution** is used when population is so small, so that the outcome of a trial has a large effect on the probability of the next outcome.

The hyper-geometric distribution is used under these conditions:

 Total number of items (population) is fixed.
Sample size (number of trials) is a portion of the population.
Probability of success changes after each trial.

The fundamental difference between hyper-geometric distribution and geometric distribution

 The difference between hyper-geometric distribution and geometric distribution is in contrast to the Bernoulli, binomial, and hyper-geometric distributions, where the number of possible values is finite. Whereas, in the geometric and negative binomial distributions, the number of "successes" is fixed, and we count the number of trials needed to obtain the desired number of "successes".

In such a sequence of trials, the geometric distribution is useful to model the number of failures before the first success. The distribution gives the probability that there are zero failures before the first success, one failure before the first success, two failures before the first success, and so on.

 **Case Study 7**

A group of 6 female managers and 19 male managers applies for an assignment. A random sample of 5 people is draw from a hat without replacement. What is the probability that 4 of the chosen manages will be a female and 1 will be a male?

$$P(X = k) = \frac{(\frac{6}{4})(\frac{19}{1})}{\frac{25}{5}}$$

Case Study 8

Suppose that a shipment contains 5 defective items and 10 non-defective items. If 7 items are selected at random without replacement. What is the probability that at least 3 defective items will be obtained?
N=15 (5 defective, 10 non defective)
N=7

$$P(0) = \frac{(\frac{5}{0})(\frac{10}{7})}{\frac{15}{7}} = 0.0176$$

$$P(1) = \frac{(\frac{5}{1})(\frac{10}{6})}{\frac{15}{7}} = 0.1731$$

$$P(2) = \frac{(\frac{5}{2})(\frac{10}{5})}{\frac{15}{7}} = 0.3917$$

$$P(X = 3) = 1 - P(x \leq 2) = 1 - [P(o) + P(1) + P(2)] = 0.4267$$

Case Study 9

A hat contain 5 green marbles and 9 blue marbles. 4 marbles are drawn randomly without replacement. Calculate each probabilities;

a. The probability of getting 3 green marbles;

$a = 5; n = 14 \; and \; r = 4$

$$P(X = 3) = \frac{C_3^5 \times C_1^9}{C_4^{14}} = \frac{(10)(9)}{(1001)} = 0.00991008$$

Self-Assessment Exercise(s)

1. What is a **Geometric Distribution**?
2. What is the difference between hyper-geometric distribution and geometric distribution?
3. What are the requirements of a hyper-geometric distribution?
4. What are the four conditions of a geometric distribution?
5. What is the difference between binomial and geometric distribution?

3.4 Summary

The Poisson distribution describes the probability to find exactly *x* events in a given length of time if the events occur independently at a constant rate while the geometric distribution is a special case of the negative binomial distribution. It deals with the number of trials required for a single success. Thus, the geometric distribution is a negative binomial **distribution** where the number of successes (r) is equal to 1. Finally, the hypergeometric distribution is a discrete probability distribution that describes the probability of k successes (random draws for which the object drawn has a specified feature) in n draws, *without* replacement, from a finite population of size N that contains exactly K objects with that feature, wherein each draw is either a success or a failure.

Unit 3 establishes that the discrete probability distribution of the number of events occurring in a given time period, given the average number of times the event occurs over that time period.

3.5 References/Further Reading/Web Resources

Robert K. and Jim Poserina, (2017),Optimal Sports Math, Statistics, and Fantasy.

Sinharay S. (2010), International Encyclopedia of Education (Third Edition).

Yates, R. D. and Goodman, D. J. (2014), Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers (2nd ed.), Hoboken, USA: Wiley, ISBN 978-0-471-45259-1

https://en.wikipedia.org/wiki/Hypergeometric_distribution#:~:text=In%20probability%20theory%20and%20statistics,that%20contains%20exactly%20objects%20with

https://en.wikipedia.org/wiki/Poisson_distribution#:~:text=In%20probability%20theory%20and%20statistics,time%20or%20space%20if%20these

MODULE 6 ESTIMATION AND HYPOTHESIS TESTING

The details of estimating population parameters and hypothesis testing from samples were explicated in units: 1, 2 and 3.

Unit 1 Estimation
Unit 2 Principle of Hypothesis Testing
Unit 3 Statistical Hypotheses' Dimensions

UNIT 1 ESTIMATION

Unit Structure

1.1 Introduction

Estimation is any of numerous procedures used to calculate the value of some parameters of a population from observations of a sample drawn from it.

1.2 Intended Learning Outcomes (ILOs)

By the end of this unit, you will be able to learn different procedures used to calculate the value of some property of a population from observations of a sample drawn from the population.

1.3 Main Content

1.3.1 Definition of terms in Estimation:

Estimator- This a rule, often expressed as a formula that tells how to calculate the value of an estimate based on the measurement obtained in a sample. Estimator is the formula or rule used in getting the unknown parameter. An estimator is particular example of a statistic, which becomes an estimate when the formula is replaced with actual observed sample values. An estimator is a statistic e.g. $\bar{X}$ is an estimator of the population mean μ and $S^2$is an estimator of the population variance σ^2 .

Estimate- An estimate is the actual value obtained from an estimator when sample is taken or the sample realization of an estimator. It can be a particular value (point estimation) or an interval value (interval estimation). This is the actual value of the estimator.

Case Study 1

If $\bar{X}$ = 55, the sample mean, then 55 is the estimate.

Estimation- Estimation involves using the sample data to get a value for the unknown parameter. This implies that in estimation we use part of the entire population to draw conclusion. It represents ways or a process of learning and determining the population parameter based on the model fitted to the data. Estimation involves approximating the value of an unknown parameter

1.3.2 Types of Estimation

Point estimation and interval estimation, and hypothesis testing are three main ways of learning about the population parameter from the sample statistic.

1.3.2.1. Point Estimation

When the estimate gotten from sample observation is used as an estimate of a population parameter is a single value then it is called point estimation. Point estimates are single values calculated from the sample. This is the process of obtaining a single numerical value from a random sample to estimate the unknown population parameter.

 **Case Study 2**

21 years is the mean age of students in this class.

Interval Estimation
This is the process of obtaining an estimate of the population parameter as being within an interval L and U, which are functions of the observed random variable. The probability that the parameter lie between the interval L and U i.e. $\rho\ [L \leq\ \theta \leq U\ ]$ where θ is the parameter, is expressed in terms of predetermined number, $1 - \alpha$, called the confidence coefficient and α is the level of significance. L is called the lower limit and U the upper limit. The interval (L,U) is called 100 $(1-\alpha)$% confidence interval.

In statistics, interval estimation is the use of sample data to calculate an interval of possible (or probable) values of an unknown population parameter, in contrast to point estimate.

1.3.2.3 Confidence Intervals gives a range of values for the parameter interval estimates are intervals within which the parameter is expected to fall, with a certain degree of confidence.
Hypothesis tests = tests for a specific value(s) of the parameter.

1.3.3 Method of Estimation

1.3.3.1 Least Square Estimation

The method of least squares is a standard approach in regression analysis to the approximate solution of over determined system, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation.

1.3.3.2 Method of Moment

Here, it involves equating the kth sample moment to the population moment.

For instance,

Sample moment is given as $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$ 3.1

In other words, method of moment involves equating the kth sample moment to the corresponding population moment and solve the resultant equation for the estimate of the parameter.

1.3.3.3 Maximum Likelihood Estimate

In Statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model with given observations, by finding the parameter values that maximize the likelihood function. MLE can be seen as a special case of the maximum a posteriori estimation (MAE) that assumes a uniform prior distribution of the parameters, or as a variant of the MAE that ignores the prior and which therefore is un-regularized.

The method of maximum likelihood corresponds to many well-known estimation methods in statistics. For example, one may be interested in the heights of adult female penguins, but be unable to measure the height of every single penguin in a population due to cost or time constraints.

1.3.4 Criteria of Estimation

1.3.3.1: Consistency

In Statistics, consistency of procedures, such as computing confidence intervals or conducting hypothesis tests, is a desired property of their behaviour as the number of items in the data set to which they are applied increases indefinitely.

Definition of Consistency: An estimator is consistent if the probability that it equals the parameter being estimated approaches 1 as $n \Longrightarrow \infty$

i.e $\text{Lim}_{n\to\infty} \rho\left[/\hat{\theta} - \mu/\alpha\varepsilon\right] \to 1$

$\text{Lim}_{n\to\infty}$ Var $\left(\hat{\theta}\right) \to 0$ 3.2

1.3.4.2 Unbiasedness

In Statistics, the bias (or bias function) of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated. An estimator or decision rule with zero bias is called unbiased. Otherwise the estimator is said to be biased.

Definition of Unbiasedness: An estimator is said to be unbiased if its expectation is equal to the parameter estimated.

 **Case Study 3:**
$E(\overline{X}) = \mu \Longrightarrow \overline{X}$ is unbiased for μ, where $\overline{X}$ is an estimator of μ.

1.3.4.3 Efficiency

In the comparison of various statistical procedures, efficiency is a measure of quality of an estimator, of an experimental design, or of a hypothesis testing procedure. Essentially, a more efficient estimator, experiment, or test needs fewer observations than a less efficient one to achieve a given performance.

Definition of Efficiency: If $\overline{X}_1$ and $\overline{X}_2$ two unbiased estimators of the population mean μ, we say the $\overline{X}_1$ is more efficient that $\overline{X}_2$ if Var $(\overline{X}_1)$ < var $(\overline{X}_2)$.

1.3.4.4: Minimum Variance

In Statistics a minimum-variance unbiased estimator (MVUE) or uniformly minimum-variance unbiased estimator (UMVUE) is an unbiased estimator that has lower variance than any other unbiased estimator for all possible values of the parameter.

1.3.4.5: Completeness

In Statistics, completeness is a property of a statistic in relation to a model for a set of observed data set. It is closely related to the idea of identifiability, but in statistical theory it is often found as a condition imposed on a sufficient statistic from which certain optimality results are derived.

1.3.4.6 Sufficiency: Definition of Sufficiency

An estimator is said to be sufficient if it gives all the necessary information, which no other estimator can provide.

SELF-ASSESSMENT EXERCISE(S)

1. Define the following terms;
   a. Estimation
   b. Estimator
   c. Estimate
   d. Estimation
2. List and explain types of estimation
3. What is confidence intervals?
4. Define the method of estimation
5. Explain the following criteria of estimation
   a. Consistency
   b. Unbiasedness
   c. Efficiency
   d. Minimum Variance
   e. Completeness
   f. Sufficiency

1.4 Summary

Unit 1 captures different procedures used to calculate the value of some property of a population from observations of a sample drawn from the population.

This unit covers estimation, which is any of numerous procedures used to calculate the value of some parameters of a population from observations of a sample drawn from it.

1.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Alreck, P.L. and Settle, R. (2003). Survey Research Handbook, 3rd edition, McGraw Hill, New York.

Barnes, S. (Ed.) (2005). News of the World, Football Annual 2005–2006, Invincible Press, London, ISBN 0-00-720582-1.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

UNIT 2 PRINCIPLE OF HYPOTHESIS TESTING

Unit Structure

2.1 Introduction

The theory, methods, and practice of testing a hypothesis by comparing it with the null hypothesis. The null hypothesis is only rejected if its probability falls below a predetermined significance level, in which case the hypothesis being tested is said to have that level of significance.

2.2 Learning Outcomes (ILOs)

To establish that a statements or assumptions about the parameters of known distribution

2.3 Main Content

2.3.1 Test of Significance

2.3.2 Statistical hypotheses: These are statements or assumptions about the parameters of known distribution which may or may not be true.

Case Study 1

The mean age of students offering STA 312 is 21 years.

2.3.2.1 Types of hypotheses

(i) Null hypothesis: This is the hypothesis being tested and it is formulated for the sole purpose of being rejected it is denoted by Ho.
Example: Ho: mean age of students offering STA 312 is 21 years.

(ii) Alternative hypothesis: This is any other hypothesis contradicting the null hypothesis and it is denoted by H_1. Example: H_1: the mean age of students offering STA 312 is more than 21years or not equal to 21years as the case may be.

Whether null or alternative hypothesis it can either be simple or composite. A simple hypothesis is one, which specifies all be values of the parameter.

 **Case Study 2:. H_o: $\pi_1 = \pi_o$**

H_o: $\pi = \pi_o$;

H_o: $\pi = \pi_1$, π_0 and π_1 are specified values of the population proportion and mean respectively. However, when all the values of the parameter are not specified completely by the hypothesis, we have a composite hypothesis.

 **Case Study 4:**

H_1: $\pi \neq \pi_o$ or H_1: $\pi > \pi_o$. In each case the parameter μ or π can assume an infinite number of values different from μ_o and πo.

 **2.3.3 Level of significance**:

Generally in testing hypothesis we make a decision to accept or reject a null hypothesis. Since the decision is based on sample information, we realize that whatever decision we make, we stand the chance of committing two distinct types of error. These errors are

(1) Type 1 error: We commit type 1 error when we reject the null hypothesis when it is correct. The probability of committing type 1 error is denoted by α and this error is usually reduced to barest minimum. The maximum probability with which we would be willing to risk a type I error is called the **level of significance** or simply the level of significance of a test. It is the maximum probability of committing type I error usually the level of significance is denoted by α.

(2) Type II error. This is the acceptance of a null hypothesis when it is false. The probability of committing type II error also referred to as the size of the type II error is denoted by β.

2.3.4 Interpretation of α

If in test of hypothesis α =0.05, this means that 5 out of every 100 cases we shall be rejecting a true null hypothesis which implies that we are 95 percent confident that we have made the correct decision.

Decision state of nature	H_o is true	H_1 is true
Reject H_o	Type 1 error	Correct Decision
Accept H_o	Correct Decision	Type II error

2.3.5 The Power of a Test

The power of a test is the probability of rejecting the null hypothesis when it is actually false. When testing hypothesis about a parameter μ, the probability of accepting H_0 is a function of μ denoted by c(μ) and is called the operating characteristic 0C curve of the test. The 0C curve describes how the probability of type II error varies with μ…

The power function of a test of statistical hypothesis is a function which yields the probability of rejecting the null hypothesis under consideration. The value of the power function at a parameter point is called the power of the test at that point.

2.3.5.1 Types of Test

Generally the null hypothesis is stated in a specified form

Case Study 4: $H_{o:}$ $\mu = \mu_o$ and the Alternative is stated in such a non-specified form such as
H_i: $\mu < \mu_o$ or H_i: $\mu > \mu_o$ or H_i: $\mu \neq \mu_o$

1. <u>One tailed test:</u> When the alternative hypothesis is stated as H_i: $\mu > \mu_o$, we have a one tailed test or one-sided test to the right. When the alternative hypothesis is stated as H_i: $\mu < \mu_o$, we have a one tailed or one-sided test to the left. In each case we are interested only in extreme values to one side of the distribution.

The critical region is such test is a region to one side of the distribution with the area equal to the level of significance

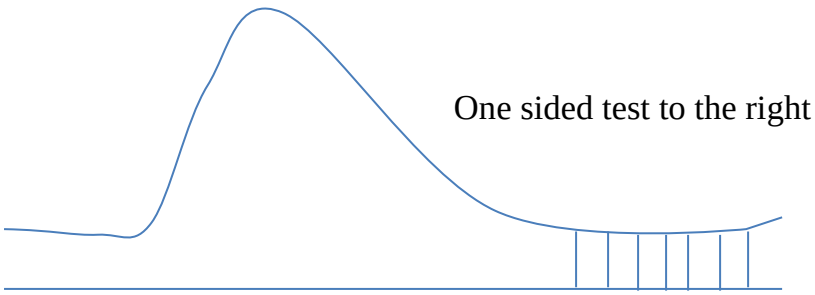
2. Two tailed or two sided test: when the alternative hypothesis is stated as H_i: $\mu \neq \mu_o$ we have a two tailed test. In this test we are interested in extreme values of the test statistic or its corresponding Z-score on both sides or tails of the distribution. In this case there are two critical regions, one to the right extreme and the other to the left extreme of the distribution. The combine areas of these two regions are α. The two regions are equal with each equal to half of α.

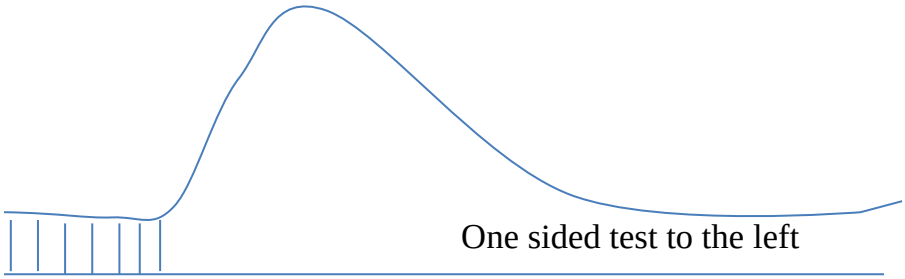
2.3.6 Critical and Acceptance Regions

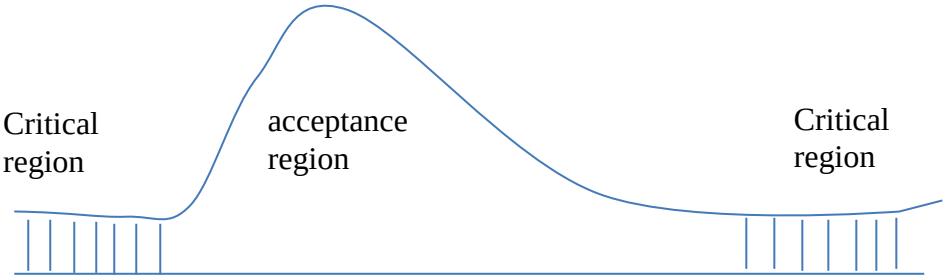
a. Critical Region: The critical region of a test is the region of the test that leads to the rejection of the null hypothesis. It is also called the rejection region or region of significance. If the value of the test statistic falls into this region, the null hypothesis is rejected.

b. Critical value: This is the value, which separates the critical region from the acceptance region, and it depends on the level of significance and in some cases on the degree of freedom of the test.

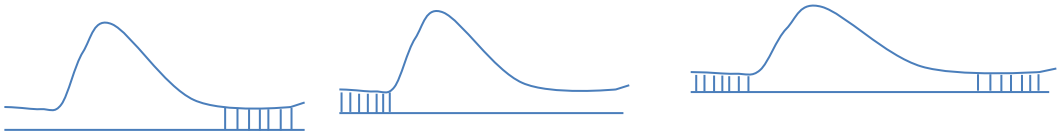
c. Acceptance Region: This is the region of the distribution that leads to the acceptance of the null hypothesis. It is also called region of non-significance.







The shaded regions are the critical regions



2.3.7 Procedure for Testing Hypothesis

The procedure for testing hypothesis consists of the following steps:

(a) Formulate the null and appropriate alternative hypotheses

(b) Choose the level of significance if it is not given

(c) State the basic assumption. These are statement about the population parameters from which the samples are drawn.

(d) Choose the test statistic which should not contain any unknown parameter and whose distribution under the null hypothesis is known.

(e) Construct the decision rule by fixing the critical region depending on the alternative hypothesis and the level of significance this involves the partitioning of the sampling distribution of the test into the rejection and the acceptance regions.

(f) Using the sample data, calculate the test statistic.

(g) Make decision to reject or accept the null hypothesis depending on whether the calculated value of the test statistic falls into the critical or acceptance region.

SELF-ASSESSMENT EXERCISE(S)

1. What is the difference between statistical hypothesis and level of significance
2. Discuss types of hypothesis, types of error, types of test
3. What do you understand by the following: power of a test, interpretation of α, critical and acceptance regions
4. Identify the difference between critical region, critical value, acceptance region including their diagrams
5. Itemize the procedure for testing hypothesis

2.4 Summary

Unit 2 handles the basic principles of hypothesis testing. It also teaches us on how to make decisions whether to reject or accept the null hypothesis depending on the calculated value of the test statistic.

This unit establishes statements or assumptions about the parameters of known distribution

2.5 References/Further Readings/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions", The American Statistician, 52(2), 119-126.

Alreck, P.L. and Settle, R. (2003). Survey Research Handbook, 3rd edition, McGraw Hill, New York.

Barnes, S. (Ed.) (2005). News of the World, Football Annual 2005–2006, Invincible Press, London, ISBN 0-00-720582-1.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

Bhattacharyya, G. K., and R. A. Johnson, (1997). Statistical Concepts and Methods, John Wiley and Sons, New York.

Moore, David S., "The Basic Practice of Statistics." Third edition. W.H. Freeman and Company. New York. 2003.

UNIT 3 STATISTICAL HYPOTHESES' DIMENSIONS

Unit Structure

3.1 Introduction

Statistical hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results. It is an act in statistics whereby an analyst tests an assumption regarding a population parameter.

3.2 Intended Learning Outcomes (ILOs)

To examine two opposing hypotheses about a population whether there is enough evidence to infer that a convinced condition is true for the entire population.

3.3 Main Content

3.3.1 Branches of Hypotheses

i. Simple Test of Hypothesis
ii. Composite Test of Hypothesis

3.3.1.1 Simple Test of Hypothesis

A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population.

It examines two opposing hypotheses about a population: **the null hypothesis and the alternative hypothesis**.

- **The null hypothesis** is the statement being tested. Usually the null hypothesis is a statement of "no effect" or "no difference".
- **The alternative hypothesis** is the statement you want to be able to conclude is true.

Based on the sample data, the test determines whether to reject the null hypothesis. You use a p-value, to make the determination. If the p-value is less than or equal to the level of significance, which is a cut-off point that you define, then you can reject the null hypothesis.

3.3.1.2 Composite Test of Hypothesis

When a set contains more than one parameter value, then the hypothesis is called a composite hypothesis, because it involves more than one model.

3.3.2 Test of Hypotheses for Small and Large Samples

3.3.2.1 Large Sample Test for one Population Mean

a) One tailed (right tailed)

$H_0: \mu \leq \mu_0$

$H_1: \mu > \mu_0$

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \rightarrow \textit{the test statistic}$$

Decision rule: For a specified α, reject H_0 if the computed test value $Z_{cal} > +Z_{tab.}$

b) One tailed (left tailed)

$H_0: \mu \geq \mu_0$

$H_1: \mu < \mu_0$

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \rightarrow \textit{the test statistic}$$

Decision rule: For a specified α, reject H_0 if the computed test value $Z_{cal} < -Z_{tab}$

c) Two tailed (right tailed)

H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \rightarrow \textit{the test statistic}$$

Decision rule: For a specified α, reject H_0 if the computed test value $Z_{cal} < -Z_{tab}$

Case Study 1:

The teachers union will like to establish that the average salary of a high school teachers in a particular state is less than $32,500. A random sample of 100 public school teachers in a particular state has a mean salary of $31.578. it is known from past history that standard deviation of the salary for the teachers in the state is $4,415. Test the union's claim at 5% level of significant.

Solution

H_0: $\mu \geq 32{,}500$

H_1: $\mu < 32{,}500$

n=100, $\bar{x} = 31{,}578$, $\sigma = 4415$

α=5%=0.05

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \rightarrow \textit{the test statistic}$$

$$= \frac{31578 - 32500}{\frac{4415}{\sqrt{100}}} = \frac{-922}{441.5} = -2.0883$$

Decision rule: For a specified value of α =0.05, reject H_0 if the test value $Z_{cal} < -Z_{tab}$ ie if -2.0883 <-1.645

Conclusion: Since Z_{cal} =-2.0883 $< -Z_{tab}$=-1.645, we reject the null hypothesis at α=0.05. Therefore, there is a sufficient evidence to reject the union's claim.

3.3.2.2 Large Sample for two Populations Mean

a) One tailed (Right tailed)

H_0: $\mu_1 \leq \mu_2$

H_1: $\mu_1 > \mu_2$

Test statistic, $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

Decision rule: For a specified value of α, reject H_0 if the computed test value, $Z_{cal} > +Z_{tab}$

b) One tailed (left tailed)

H_0: $\mu_0 \geq \mu_2$

H_1: $\mu_1 < \mu_2$

Test statistic, $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

Decision rule: For a specified value of α, reject H_0 if the computed test value, $Z_{cal} < -Z_{tab}$

c) Two tailed

H_0: $\mu_0 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

Test statistic, $Z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

Decision rule: For a specified value of α, reject H_0 if the computed test value, $Z_{cal} > +Z_{tab}$ or $Z_{cal} < -Z_{tab}$ where; $Z_{tab} = Z_{\alpha/2}$

3.3.3 Population Proportion

3.3.3.1 Large Sample Test for One Proportion

a) One tailed (right tailed)

H_0: $p \leq p_0$

H_1: $p > p_0$

The test statistic: $Z = \dfrac{x - np_0}{\sqrt{np_0(1 - p_0)}}$

Decision rule for right tail

For a specified α, reject the null hypothesis if the computed statistic value Z is greater than Z_α (*Z tabulated*) *ie.* $|Z_{cal}| > |Z_{tab}|$

b) One tailed (left tailed)

H_0: $p \geq p_0$

H_1: $p < p_0$

The test statistic: $Z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$

Decision rule for left tail

For a specified α, reject the null hypothesis if the computed statistic value Z is less than $Z_\alpha (Z\ tabulated)\ ie. |Z_{cal}| < |Z_{tab}|$

c) Two tailed

H_0: $p = p_0$

H_1: $p \neq p_0$

The test statistic: $Z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$

Decision rule

For a specified α, reject the null hypothesis if the computed statistic value Z is less than $Z_{\frac{\alpha}{2}} ie\ Z_{cal} < -Z_{tab}$ or if it is greater than $Z_{\frac{\alpha}{2}} ie\ Z_{cal} > Z_{\frac{\alpha}{2}}$

Case Study 2

Your teacher claim that 60% of the Nigerian males are married. You fill that the proportion is higher. In a random sample of 100 Nigerian males, 65 of them are married. Test your teacher's claim at 5% level of significant.

Solution

Here:

P_0 =60% =0.06, x=65

H_0= $p \leq 0.6$ n=100

H_1: p>0.6

α=5%=5/100=0.05

$$Z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{65 - 100(0.6)}{\sqrt{100(0.6)(1-0.6)}}$$

$$\frac{65-60}{\sqrt{60(0.4)}} = \frac{5}{\sqrt{24}}$$

$$\frac{5}{4.8990} = 1.0206$$

$$Z_{tab} = Z_\alpha = 0.5 - 0.05 = 0.45$$

From the table; $Z_\alpha = Z_{0.45} = 1.65$

Decision rule: For a specified value of α =0.05, we reject the H_0 if the Z_{cal} =1.0206 <Z_{tab}=1.65, we do reject the null hypothesis at α=0.05, showing that there is significant claim to reject your teacher's claim.

3.3.3.2 Large Sample Test between Two Population Proportions

1) One tailed (left tailed)
   H_0: $P_1 \geq P_2$
   H_1: $P_1 < P_2$
   Test statistic,

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$???

Decision rule: For a specified value of α, reject H_0 if the computed test value,-Z_{cal}<-Ztab

2) Two tailed
   H_0: $P_1 \neq P_2$
   H_1: $P_1 \neq P_2$
   Test statistic,

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Decision rule: For a specified value of α, reject H_0 if the computed test value, Z_{cal}>+Z=tab $\Rightarrow Z_{tab} > +Z_{\frac{\alpha}{2}}$ *or* $Z_{cal} < -Z_{tab} \Rightarrow Z_{cal} < -Z_{\alpha/2}$

where

$$\hat{P} = \frac{x_1 + x_2}{n_1 + n_2}$$

3.4 Hypothesis Testing in Small Sample

a) One-Tailed (Right Tailed)
H_0: $\mu_1 \leq \mu_0$
H_1: $\mu_1 > \mu_0$
Test statistic,

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Decision rule: For a specified value of α, reject H_0 if the computed test value,
tcal > t_α (n-1)tab.

SELF-ASSESSMENT EXERCISE(S)

1. Define the hypothesis test and explain the two types of hypothesis testing.
2. What is the difference between the null hypothesis and the alternative hypothesis?
3. The Dean of students of a community college claims that the average distance that community students travelled to the campus is 32 km. the community students feel otherwise. The sample of 64 students were randomly collected and yielded the mean 33km and standard deviation of 5km. test the Dean's claim at 5% level of significant.
4. A random sample of size n_1=36 is selected from a normal population distributed with σ_1=4, $\bar{x} = 75$. A second random sample of size n_2 =25 is also selected from a different normal distribution with $\sigma_2 = 6$ *and* $\bar{x}_2$=25. Is there any significant difference between the population mean at 5% level of significance?
5. A preacher will like to establish that of people who pray less than 80% prayed for world peace. In a random sample of 110 persons. 77 of them said that when they pray, that they prayed for world peace. Test at 10% level of significance.

3.5 Summary

Unit 3 observes two opposing hypotheses about a population whether there is enough evidence to infer that a convinced condition is true for the entire population or not.

This unit examines two opposing hypotheses about a population whether there is enough evidence to infer that a convinced condition is true for the entire population.

3.6 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions", The American Statistician, 52(2), 119-126.

Alreck, P.L. and Settle, R. (2003). Survey Research Handbook, 3rd edition, McGraw Hill, New York.

Barnes, S. (Ed.) (2005). News of the World, Football Annual 2005–2006, Invincible Press, London, ISBN 0-00-720582-1.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

Bhattacharyya, G. K., and R. A. Johnson, (1997). Statistical Concepts and Methods, John Wiley and Sons, New York.

Moore, David S., "The Basic Practice of Statistics." Third edition. W.H. Freeman and Company. New York. 2003.

MODULE 7 PROGRESSIVE STATISTICAL METHODS

Module 7 looks in progressive statistical methods in units: 1 and 2.

Unit 1 Introduction to nonparametric (Methods and test based on Runs)
Unit 2 Fundamentals of Index Number

UNIT 1 INTRODUCTION TO NONPARAMETRIC METHODS AND TEST BASED ON RUNS

Unit Structure

1.1 Introduction

Nonparametric statistics are not the solution to every data analysis problem. Some distributional assumptions are required for nonparametric procedures, and data that fail to meet parametric assumptions may also fail to meet nonparametric assumptions. Many nonparametric tests have less power than the corresponding parametric tests. Because power should never be given up unless absolutely necessary, nonparametric methods should not be used when parametric methods are appropriate. Like many parametric procedures, many nonparametric procedures, cannot be used to test hypotheses about population that consists of nominal data. But most of the nonparametric procedures we will be discussing are based on ranks. Because nominal data have no true numerical meaning, it does not make sense to rank them. Nonparametric procedures based on ranks can be used to test hypotheses about populations that consist of ordinal, interval or ratio data.

1.2 Intended Learning Outcomes (ILOs)

To use nonparametric statistical methods when data do not satisfy the distributional assumptions required by parametric procedures.

1.3 Main Content

1.3.1 Introduction to Parametric Methods

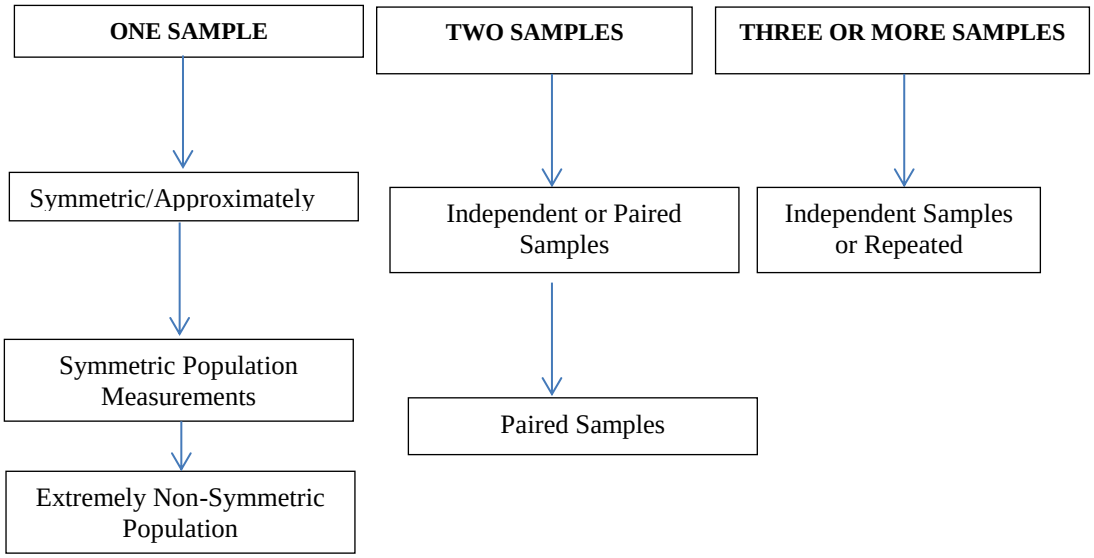
Nonparametric statistics refer to a statistical method wherein the data is not required to fit a normal distribution. Nonparametric statistics uses data that is often ordinal, meaning it does not rely on numbers, but rather a ranking or order of sorts.

In other words, nonparametric tests are statistical procedures that can be used to test hypotheses when no assumptions regarding parameters or population distribution are possible. The distributional assumptions required for nonparametric procedures are usually less specific than those required for parametric procedures; (Omekara and Acha, 2017).

The Wilcoxon signed rank test and the sign test are used to test hypothesis about population median and population median differences. The Mann-Whitney test is used to test the hypothesis that two populations are identical, and the Kruskal-Wallis test is used to test the hypothesis that three or more populations are identical. When samples of repeated measurements are obtained, the Friedman test can be used to test the hypothesis that all possible rankings of the observations from any subject are equally likely.

1.3.2 A Flow Chart for Non-Parametric Methods

A flow chart showing which of the nonparametric tests might be appropriate for particular data sets are shown below.



Advantages and Disadvantages of using Parametric rather than Nonparametric Model

A parametric model is, typically based on very specific assumption about the form of the underlying probability distribution. If these assumption are correct, then it is possible to use the data in a high efficient manner, and also possible to assess the accuracy of the inference procedure.

However, as mentioned earlier, the danger is that, if the assumptions are wrong, the conclusions may be completely irrelevant to the real problem.

Nonparametric models based on weaker, more general assumptions are applicable to a much larger class of problems than the narrow parametric models, but their very generality means that methods based on them are less sharp, and often less efficient.

1.3.3 Test Based on Runs

Definition: A run is a sequence of observations bounded by observations of different type. The number of the observations in the run is called the **Length of the run.**

 **Case Study 1**

Consider the local and international calls: TTTLL TTTT LLLL TT LLL

Solution 1
We have six runs.

 **Case Study 2**

Given binary digits to denote the number of males and females respectively;

00011111 0001111 000011001011

Solution 2
We have ten runs (10).
In a sequence of two types of observations, like the one we have above, the total number of runs can be used as a measure of the randomness of the sequence; too many runs may indicate that each observation tend to follow and be followed by an observation of the other type.
The total number of runs may be used to test the null hypothesis H_o that two independent random samples came from population with identical distribution functions.

1.3.4 Distribution of Runs (Under the hypothesis of Randomness)

Consider ''runs'' of two types of events zeros and ones.
Let r_0 and r_1 be respectively, the number of runs of zeros and ones. Then $R = r_0 + r_1$ is the total number of runs in the sequence.
We also have n_0 zeros and n_1 ones in the sequence. We want to obtain the joint pdf of r_0 and r_1.

Define $C(r_0, r_1) = \begin{cases} 2 \text{ if } r_0 = r_1 \\ 1 \text{ if } r_0 \neq r_1 \end{cases}$

The total number of ways of arranging a set of n_0 zeros and n_1 ones into $r_0 - 1$ runs of zeros and $r_1 -$ 1 runs of ones is $\binom{n_0-1}{r_0-1}\binom{n_1-1}{r_1-1}\ C(r_0 r_1)$

If the observations are random, we can think of the total number of ways of partitioning $n = n_0 + n_1$ observation into n_o and n_1 which is

$$\binom{n}{n_0} = \binom{n}{n_1} \text{ Therefore, the joint pdf of } r_0 \text{ and } r_1 \text{ is } f(r_0 r_1)$$

$$= \frac{\binom{n_0-1}{r_0-1}\binom{n_1-1}{r_1-1}}{\binom{n}{n_o}}\ C(r_0 r_1)$$

If $r_0 = r_1$the joint pdf is $f(r_0 r_1) = 2\frac{\binom{n_0-1}{r_0-1}\binom{n_1-1}{r_1-1}}{\binom{n}{n_o}}$

If the modules of r_0, r_1 that is if $[r_0 - r_1] = 1$

$$f(r_0 r_1) = \frac{\binom{n_0-1}{r_0-1}\binom{n_1-1}{r_1-1}}{\binom{n}{n_o}}$$

Find out if it can easily be shown that the marginal pdf of r_0 is $f(r_0) = \frac{\binom{n_0-1}{r_0-1}\binom{n_1-1}{r_1-1}}{\binom{n}{n_o}}$, where $r_0 = 1, 2, \ldots . n_0$

Recall that R = # number of runs in the sequence.

$R = r_0 + r_1$

If $r_0 = r_1 = r$, then R =2r. So the probability of the 2r runs is

$$f(r, r) = P(R = 2r) = 2\frac{\binom{n_0-1}{r_0-1}\binom{n_1-1}{r_1-1}}{\binom{n}{n_o}}$$

We want to find the probability of R=2r + 1. R will be equal to 2r +1 if $r_0 = r$ and $r_1 = r + 1$ or $r_0 = r + 1$ and $r_1 = r$. This is so because if we consider some sequence of zeros and ones we may observe in some sequence that $r_0 = r_1$ and in some that $r_0 = r_0 + 1$.

Case Study 4;

$$r_0 = 3, r_1 = 3 \rightarrow r_0 = r_n \text{ i.e } r_0 = 3 \; r_1 = 4$$

$$\therefore p[R = 2r + 1] = \frac{\binom{n_o-1}{r-1}\binom{n_1-1}{r-1} + \binom{n_0-1}{r}\binom{n_1-1}{r-1}}{\binom{n_0+n_1}{n_0}}$$

The distribution of R can also be calculated using t μ and σ^2. If the critical region approximated with large samples n_0 and n_1, by normal distribution with mean;

$\mu = E(R) = \frac{2nn_1}{n_0+1} + 1$ and variance, $\sigma^2 = \frac{(\mu-1)(\mu-2)}{n_0+n_1-1}$

Case Study 5:

The following given pattern of incoming trunk (T) and local (L) calls coming into a switchboard, TTT LLLL TT LLLLL TTTL. Use the information to test whether the pattern of incoming calls are random and was (T) the type trunk or local (L) of call.

Solution 5

We want to test the hypothesis.

H_o : The pattern of incoming calls is random

H_1: The pattern of incoming calls is not random

$$R = 6, n_T = 8, n_L = 10$$

From Wolf4wiz table we obtain;

$W_0.025 = 6,\ W_0.975 = 14 \rightarrow \alpha = 0.05$ and it is a two tailed test

Conclusion

Since $6 \leq R \leq 14$, we accept H_o at 5% level of significance.

1.3.5 Importance of Runs Test

Test for Randomness:

A test for randomness is a non-parametric test that has wide area of application in the field of Statistics. In Statistics, the behaviour of certain distribution may be expected to assume randomness in order not to violate some basic fundamental principles defining the distributions under study. This randomness in a distribution can be detected based on the number of runs in a sequence of distribution. In a sequence of A_s and B_s, a run is defined as maximum sequence of like elements.

Case Study 6;

Given the sequence A B B A B B B A A A then number of runs r=5.

n_A (n_1) =5 and n_B (n_B) =4

The number r of runs in a sequence of n_1 and n_2 can be used as a test statistic to test for randomness and non-randomness. The null and alternative hypotheses for the test are;

H_0: the sequence of A_s and B_s have been generated by a random process.

H_1: the sequence of A_s and B_s have not been generated by a random process.

Rejection region: one tailed test. $r \geq r_1 \ or \ r \leq r_2$

Two tailed test $r<r_1$ and $r>r_2$ – rejection region

$r \leq r_1 \ and \ r \leq r_2$ - acceptance region

When n_1 and n_2 are both large i.e, $n_1 > 10$, $n_2 > 10$, then we conduct the run test by using the formula for the standard normal test statistic

$$z = \frac{r - \mu_r}{\delta_r}; \quad \mu_r = \frac{2n_1 n_2}{n_1 + n_2} + 1 \ ; \quad \delta_r = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

SELF-ASSESSMENT EXERCISE(S)

1. State the differences between parametric or nonparametric statistical methods
2. Use a flow chart to show which of the nonparametric tests might be appropriate for particular data sets.
3. What are the advantages, or disadvantages of using a parametric rather than a nonparametric model?
4. Define a run and how many do we have in the local and international calls pattern? TTTLTTTLLTTLLLTTTLLTTTT LLLTTTLLLLTL TTLLL
5. Given binary digits to denote the number of males and females respectively, how many runs do we have: 11100011111011000111100100110010 11

1.4 Summary

Apart from establishing whether a particular dataset is discrete or continuous distribution, it is pertinent to know whether to apply parametric or nonparametric statistical methods on a particular data set. Many statistical procedures are based on specific assumptions about the distribution of the population. Statistical methods that require specific distributional assumptions are called parametric statistics.

This unit discussed nonparametric statistical methods when data do not satisfy the distributional assumptions required by parametric procedures.

1.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Applied linear Regression Weisberg 2005, 3rd Wiley.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

Bhattacharyya, G. K., and R. A. Johnson, (1997). Statistical Concepts and Methods, John Wiley and Sons, New York.

Moore, David S., "The Basic Practice of Statistics." Third edition. W.H. Freeman and Company. New York. 2003.

UNIT 2 FUNDAMENTALS OF INDEX NUMBER

Unit Structure

2.1 Introduction

Index number is a technique of measuring changes in a variable or group of variables with respect to time, geographical location or other characteristics.

2.2 Intended Learning Outcomes (ILOs)

To measure changes in a variable or group of variables.

2.3 Main Content

The consumer price index (CPI) prepared by the Nigeria Bureau of Statistics is an example of a price index while industrial production is an example of a quantity index.

2.3.1 Types of Index

When the index no. represents a comparison for an individual product or commodity, it is called simple index number. In contrast, when the index number has been constructed for a group of items or commodity, it is an aggregate index number or composite index number.

2.3.2 Construction of Index Numbers

- Price index number
- Quantity index number
- Value index number

2.3.3 Construction of Simple Price Indexes/Price Relatives

The general formula for a simple price index or price relative is

$$I_p = \frac{P_n}{P_o} X \frac{100}{1} \qquad 3.1$$

where,

P_n indicates price in given period and P_o indicates the price in the base year.

Case Study1

Referring to table 3.1 determine the simple price indexes for 1976, for the commodity, Milk using 1970 as the base year.

Table 3.1 Price and consumption of three commodities in a particular Area, 1970 and 1976.

Commodity	Unit	Average	Prices	Per capita Per month	
	Quotation	1970 P_o	1976 P_n	1970 q_o	1976, q_n
Milk	Quart	0.30	0.38	30	35
Bread	1 1b loaf	0.25	0.35	3.8	3.7
Eggs	Dozen	0.60	0.90	1.5	1.0

Solution 1:

For milk = $\frac{0.38}{0.30} X 100 = 126.67,$

2.3.4 Construction of Simple Quantity Indexes/Quantity Relative

 The general formula for simple price index or quantity relative is

$$\frac{q_n}{q_o} X \frac{100}{1} \qquad 3.2$$

where q_n indicates the quantity of an item produced or sold in the given period and q_o indicates the quantity in the base year.

 **Case Study 2:**
Referring for table 3.1. Determine simple quantity indexes for the three commodities for 1976, using 1970 as the base year.

Solution 2:

For Milk = $\frac{35}{30} X 100 = 116.7$

For Bread = $\frac{3.7}{3.8} X 100 = 97.37$

For Eggs = $\frac{1.0}{1.5} X 100 = 66.67$

Finally, the value of a commodity in a designated period is equal to the price of the commodity multiplied by the quantity produced (or sold).

2.3.5 Construction of Simple Price Indexes/ Value Relative

 Therefore $P_n q_n$ indicates the value of a commodity in the given period and the $P_o q_o$ indicates the value of the commodity in the base period. The general formula for a simple price index or value relative is

$$I = \frac{P_n q_n}{P_o q_o} X \frac{100}{1} \qquad 3.3$$

 **Case Study 3:**
Compute the simple value relative for 1976 for the commodity - Milk in table 22.1 using 1970 as the base year.

Solution 3:

For milk = $I = \frac{P_n q_n}{P_o q_0} X \frac{100}{1} = \frac{0.38 X 3.5}{0.30 X 30} X \frac{100}{1} = 147.8$

 **Aggregate Price Index**:

To obtain an aggregate price index, the prices of the several items or commodities could easily be summed for the given period and for the base year and then compared. Such an index would be an un-weighted aggregate price index.

 **Construction of an Un-weighted Aggregate Price Index**

An un-weighted index is generally very useful because the implicit weight of each item in the index depends on the unit upon which the prices are based.

Example: If the price of milk is reported per gallon as contrasted to then the price will make a much greater contribution to an un-weight price index for a group of commodities.

Weighted Aggregate

Because of the difficulty described above, aggregate price indexes are generally weighted according to the quantities q of the commodities. The question at which period of quantities should be used which serves as the basis for different aggregate price relatives.

Two Method of Constructing the Weighted Aggregate Price Index (WAPI)

One of the more popular aggregate price indexes:

1. **Laspeyres Index**: In which the prices are weighted by the quantities associated with the base year before being summed.

$$I(L) = \frac{\sum P_n q_o}{\sum P_o q_o} X \frac{100}{1} \qquad 3.4$$

 **Case Study 4**

Compute Laspeyres aggregate price index for 1976 for three commodities in table 3.1 using 1970 as the base year.

Table 3.2 Worksheet for the calculations of Laspeyres index for the data in table 3.1

Commodity	P_nq_o	P_oq_n
Milk	11.40	9.00
Bread	1.33	0.95
Egg	1.35	0.90
Total	$\sum P_n q_o = \&14.08$	$\sum P_o q_o = \&10.85$

Solution 4:
Referring to table 3.2 the index is determined as follows

Paasche's Index
From 3.3, we can calculate the Paasches index as follows

$$I(P) = \frac{\sum P_n q_n}{\sum P_o qn} X \frac{100}{1} \qquad 3.5$$

 **Case Study 5**
Compute Paasches aggregate price indexes for 1976 for Milk in table 3.1 using 1970 as the base year.

Table 3.3 Worksheet for the calculate of Paasches' index for the data in table 3.1

Commodity	P_nq_n	P_oq_o
Milk	13.5	10.50
Bread	1.30	0.93
Egg	0.90	0.60
Total	$\sum P_n q_n = 15.50$	$\sum P_o q_n = 12.03$

Solution .5
For Milk = $\frac{13.5}{12.03} X \frac{100}{1} = 128.6$

 **Case Study 6**
Compute the price index by the weighted average of price relatives method for the 3 commodities from table 21.1. Using 1970 as the base year.

$$I_p = \frac{\sum P_o q_o (P_n / P_o X 100}{\sum P_o q_o} \qquad 3.6$$

With reference to Table 3.4: Worksheet for the computation of the weighted average of price relatives for the data in Table 3.1.

Solution.6:
Table 3.4: The weighted average of price relatives for the data in Table 3.1.

Commodity	Price relative P_n/P_oX100	Value weighted P_oq_o	Weighted relative $(P_oq_o)(P_n/P_oX100)$
Mil	126.67	9	1140.03
Bread	140	0.95	133
Eggs	150	0.9	135
Total	416.67	10.85	1408.03

Uses of Index Number

- It can be used to measure one's real income. The quantities of goods and services that can be purchased by a fixed amount of money changes as the price decreases.
- It is used in measuring the inflation rate.
- Price comparison below locations can also be measured using index numbers.
- It is used in measuring changes in quantity.

Limitations of Index Numbers

- Selection of base period which has to be updated from time to time.
- Collection of data: Here a survey needs to be carried out which involve time and money.
- Choice of representative class (items): It may be practically difficult to consider all possible outcomes of an event. To achieve this, a list of all possible items are compiled, grouped into sections as representative items
- Changes due to time. For e.g. 1950 to 1980 are to be compared. Some items available in 1980 were not available in 1950 which make the comparison very difficult.
- Choice of weight: Here to assign the appropriate weight to the item to the items are difficult.

SELF-ASSESSMENT EXERCISE(S)

1. What is index number
2. List and explain types of indexes
3. Discuss the three major ways of constructing index numbers
4. Define aggregate price index:
5. Differentiate between an un-weighted aggregate price index and weighted aggregate

2.4 Summary

This unit estimates the percentage relative number by which a measurement in a given period is expressed as a ratio to the measurements which involves the quantity, price and value.

In summary, an index number is a percentage relative by which a measurement in a given period is expressed as a ratio to the measurements which involves the quantity, price and value.

2.5 References/Further Reading/Web Resources

Acha C. K. (2018). Fundamentals of Statistics. ISBN: 978-2832-50-2. Daves Publishers, Uyo.

Applied linear Regression Weisberg 2005, 3rd Wiley.

Berenson, M.L., Levine, D.M., and Krehbiel, T.C. (2006). Basic Business Statistics, 10th edition, Pearson, Prentice Hall, New Jersey, ISBN 0-13-196869-6.

Bhattacharyya, G. K., and R. A. Johnson, (1997). Statistical Concepts and Methods, John Wiley and Sons, New York.

Moore, David S., "The Basic Practice of Statistics." Third edition. W.H. Freeman and Company. New York. 2003