MAKERERE UNIVERSITY SCHOOL OF PUBLIC HEALTH

MPH 7103: APPLIED BIOSTATISTICS AND INFORMATICS I

Coordinating Department
Epidemiology and Biostatistics

Master of Public Health
Distance Education Programme

COURSE MATERIALS

4th Edition
(May 2019)

Contributors

Assoc. Prof. Fredrick Makumbi

Dr. Simon Kasasa

Assoc. Prof. David Guwatudde

Assoc. Prof. Nazarius Mbona Tumwesigye

Assoc. Prof. Noah Kiwanuka

Dr. Mayega Roy William

Mary Nakafeero

Instructional Design/Editing by

Dr. Mayega Roy William, MBChB, MPH, PhD.

Makerere University School of Public Health

Edition 4

(May 2019)

TABLE OF CONTENTS

Makerere University School of Public Health
Master of Public Health – Distance Education Programme

1.0 ABOUT THIS COURSE

1.1 COURSE TITTLE: Applied Biostatistics and Informatics I

Course Code: **MPH 7103**
Credit Units: **3**
Course Coordinator: **Dr. Frederick Edward Makumbi**
Email: fmakumbi@musph.ac.ug
Tel: +256-41-545001

Instructional Designer: Dr. Mayega Roy William
Email: rmayega@musph.ac.ug
Tel: +256-772-412455

1.2 GENERAL OUTLINE OF THE COURSE

This course will cover the following areas:

- Unit 1: Basic Statistical Concepts
- Unit 2: Descriptive Statistics
- Unit 3: Probability and Probability Distributions
- Unit 4: Inferential Statistics Part A: Estimation and Hypothesis testing
- Unit 5: Inferential Statistics Part B

1.3 INTRODUCTION

Welcome message from the Course Coordinator: As coordinator for Applied Biostatistics and Informatics I, I welcome you to the exciting field of Biostatistics. As you will eventually learn, Statistics cuts across all disciplines including Public Health, and in addition to Epidemiology, is one of the core courses in public health.

About this course: This course has both theoretical and practical sessions. It is divided into four units as outlined above. In the practice of epidemiology, public health and many other fields, you will often use statistical methods, including description of populations using parameters derived from data, estimation and hypothesis testing. Biostatistics methods are also used to infer associations observed in samples, to entire populations – As you recall from the previous module, this is the basis for the **epidemiological approach**.

I wish to extend my sincere thanks to all the facilitators of this course for their time and dedication during the development of this course. Thank you to: Dr. Simon Kasasa, Dr. David Guwatudde, Dr. Noah Kiwanuka, Dr. Nazarius Mbona Tumwesigye, Dr. Mayega Roy William and Miss Mary Nakafeero. In case of any problems related to this course, feel free to contact me on the address indicated above.

The Course Coordinator

Assoc. Prof. Fredrick Makumbi

Assoc. Prof. Fredrick Edward Makumbi is a Senior Lecturer in the Department of Epidemiology and Biostatistics. He holds a Bachelor's Degree in Biostatistics, an MHS and a PhD in Epidemiology & Population studies. He is specialized in statistics and infectious diseases, and is widely involved in HIV/AIDS related research. He is also the PI for the Family Health and Wealth Study.

His areas of interest are: Epidemiology, the impact of HIV/AIDs on population dynamics, and HIV prevention methods (including male circumcision, and ART treatment impact of HIV incidence)

The Instructional Designer: Roy William Mayega is coordinating the improvement and enhancement of the learning materials. Specific issues related to the content and design of the materials may be routed to him on the address de_materials@musph.ac.ug.

1.4 COURSE AIMS AND INSTRUCTIONAL GOALS

1.4.1 Aim

By the end of the course, the student should be able to understand and apply the foundational/basic concepts of statistics as applied in public health. The student should therefore be able to use statistical techniques to summarize, analyse, interpret and present data for public health.

1.4.2 Instructional Goals

By the end of this course, the MPH student should demonstrate the competency to:

1. Identify and use appropriate statistical models based on the outcome variable.
2. Interpret statistical output from different models.
3. Describe the relationship between different variables (outcome, exposure and other independent) accounting for confounding and interaction terms.
4. Write a statistical report (Including data description, analysis methods and outputs with their respective interpretations).

1.5 TIME FRAME

You should devote at least 60 hours of study to this module. The module should be spread out to at least one month of study.

1.6 INSTRUCTIONAL MEDIA AND TECHNIQUES

1. **Orientation Session:** Brief instructor led face-to-face orientation sessions will be delivered at the School of Public Health at the beginning of the course. The purpose of the sessions is to give you an overview of what you are expected to read about, and guide you on the resources available.
2. **Print material:** You will be provided with hand-outs of print materials (MUSPH Distance Education Resource Kit) and other materials. You will also be provided with printed case studies where necessary. There is an additional resources kit containing a collection of hand-outs and readers that have been sorted for you. You will be expected to acquire this at your own cost. We highly recommend that you get yourself a copy of this kit. Copies are available at the DE Secretariat at the going rate for photocopying services in the school. You need to register with the secretariat to obtain a copy.

3. **E-mail:** E-mail shall be the main means of communication for submission of assignments, announcements and consultation with faculty. Please ensure that you obtain a reliable e-mail address and register it with the Programme Administrative Secretary. In case you change your e-mail, please promptly notify the same officer.
4. **E-Learning Platforms and Tools:** Learning Management tools shall be used to conduct online discussion forums and chatting. You will be informed in due course about the site to be used for hosting these interactive activities.
5. **Textbooks, internet and independent study:** You are required to search for the relevant references and acquire the core recommended readings for each course. There are also many internet sources from which you can obtain information.
6. **Activities, examples and exercises:** Please note that the biostatistics course involves a lot of calculations. It is important, especially, that you do all the exercises and examples given, so that you internalize the application of the statistical methods they convey.

Note to the Reader!

The MPH Distance Education Programme is committed to providing the best possible distance learning environment for you. As such, the Programme has invested in the development of high quality user-friendly instructional materials that will better facilitate students' learning. A full time Instructional designer/editor has been appointed for this purpose. The materials shall mainly be available in print form. CD-ROMs, interactive media, e-learning platforms and any other viable resources may from time to time be used by the Instructors to reach out to you. Assignments will be sent to your e-mail box when you have registered at the beginning of the semester. There is also an Additional Resources kit that contains additional readers, worked examples and exercises. You should endeavour to acquire a set of these readers during the face-to-face session.

1.7 MODE OF EVALUATION

1.7.1 Progressive Assessment – Self Assessment

Self-evaluation: You will be required to attempt a set of exercises and questions at the end of each presentation for your own assessment. Assignments and activities in the course are for you to test yourself and evaluate your performance. They will enable you gauge your understanding of the content. Answers to these questions may be availed in the Additional Resources section of this document. You will also be required from time to time to contribute to specific discussion points that have been set up for analysis on the discussion forums.

Participation: We may also request you to contribute meaningfully to the topics put up on the discussion boards and your contribution may be graded. In all, your active participation in the discussion boards and the posted assignments may contribute 10 to 20% of the final mark.

1.7.2 Progressive Assessment – Hand-in Assignments/Tests

In line with the University regulations, you will be evaluated in two segments; progressive assessment, which accounts for 30% of the total mark and the end of semester university examination which accounts for the remaining 70%. Apart from the self-assessment exercises, the institute shall require you to hand in one or more assignments for marking. These assignments may be in form of quizzes, structured questionnaires, long or short answer questions, term papers or project reports, to be forwarded on-line. An assignment for assessment will be indicated and will either be given to you at the time of the face-to-face, or forwarded to you by registered mail or internet. Please pay close attention to the deadlines for handing in the assignments. Progressive assessment will account for 20 to 30% of the overall course score.

TAKE NOTE THAT THE PROGRESSIVE ASSESSMENT IS A PRE-REQUISITE FOR THE END OF SEMESTER EXAMINATION.

1.7.3 University Exam

You will sit for a final course examination at the end of the semester to be held at the School of Public Health. You must make arrangements to travel to the Institute for this examination once the date has been communicated to you. This may contribute 70% of the final mark

1.8. EVALUATION OF THE INSTRUCTION PROCESS

1.8.1 PRE AND POST EVALUATION

NOTE: Before using these materials, you are kindly requested to fill the ***Evaluation Questionnaire*** for this semester and send it to the **Instructional Designer**. This questionnaire is not a test, but it will enable us to measure your expectations from the instruction process and to gauge how much you will benefit from the instruction materials for this semester, as well as inform us the extent to which the materials given in the previous semester assisted you. The questionnaire has two parts: **A Post Evaluation** of the previous materials and a **Pre-evaluation** of the materials that you expect in the new semester. In the post evaluation section, you are requested to evaluate the materials that you received in the last semester. In the pre-evaluation section, you are expected to inform us of the key competencies and qualities you expect from the new materials. Your responses will be compared with responses in the ***Post evaluation section*** *at the beginning of the next semester.* The information generated will be used in an iterative process designed to improve the materials. This evaluation may be administered during the face-to-face session, before the materials are dispatched. In the event that it is not delivered then, you can access it in your Additional Resources Folder. In case you do not fill it during the Face-to-face sessions but fill it later, you are requested to send the completed questionnaire to the Instructional Designer at the e-mail address: de_materials@musph.ac.ug.

1.8.2 PRETEST

NOTE: It is important that before you read these materials, you complete a **Pre-test**. The purpose of this test is to make a baseline assessment of what current knowledge you have. It is also an important guide to what areas you need to emphasise in your reading. For some courses, where the Course Coordinator considers it a requirement, the test may be administered at the time of face-to-face, during the introduction to this course. Otherwise for most courses, it is strictly **optional**, but you are encouraged to take it prior to your reading. The test is also contained in the **Additional Resources Folder**.

2.0 COURSE MATERIALS

2.1 Unit 1: BASIC STATISTICAL CONCEPTS

2.1.1 Introduction to the Unit

Biostatistics refers to the application of statistics in the health sciences. In this unit, you will learn the definition and meaning of biostatistics plus its role in the health sciences. You will learn some of the basic concepts and terminology used in biostatistics, measurement scales and the distinction between rates, ratios and proportions. The unit will provide a background on important sources of statistical data in everyday life. This data will be useful in planning as well as the implementation of health related interventions.

2.1.2 Unit Outline

This unit on Basic Statistical Concepts will cover the following topics:

1. Introduction: the role of biostatistics in health sciences
2. Some basic concepts in statistics
3. Sources of data
4. Measurements and measurement scales
5. Ratios, Rates and Proportions

2.1.3 Instructional goal

The MPH student should be able to describe the basic statistical concepts and relate them to applications in bio-statistics and epidemiology.

2.1.4 Unit Objectives

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

a. Define biostatistics and state its importance in health sciences
b. Explain the common concepts and terminologies used in statistics
c. Describe the different measurement scales used in statistics
d. Differentiate between a Rate, a Ratio and a Proportion

ESSENTIAL READINGS

- Wayne W Daniel (1998) Biostatistics: A Foundation for Analysis in the Health Sciences. John Wiley & Sons, Inc. 7th Edition. Chapter 1, pages 1-14

- Practical statistics for Medical Research. Douglas G Altman

2.1.5 Time Frame

1 WEEK

2.1.6 Content

Lesson 1: Introduction to Biostatistics

Introduction: It is important to note that the tools of biostatistics are used in many fields including health, education, engineering, transport, economics, psychology and many others. This lesson will give a background on the importance of statistics in health, and in particular, the application of statistics in epidemiology and public health.

Lesson Topics: The following topics will be covered: -

a. The role of biostatistics in health sciences
b. Some basic concepts in biostatistics
c. Sources of data

Lesson Objectives:

By the end of this lesson, the MPH student should be able to:

1. Illustrate the role of bio-statistics in the health sciences
2. Describe the basic concepts in statistics
3. Examine health systems to distinguish the possible sources of statistical data

a. Introduction: the role of biostatistics in health sciences

Background: Before we explore the definition and the role played by statistics in health sciences, let us first define the term STATISTICS in general. There are many ways that we can broadly define statistics. Statistics is a field of study that deals with:

- Collection
- Organization
- Summarizing
- Analysis of numerical data
- Presentation

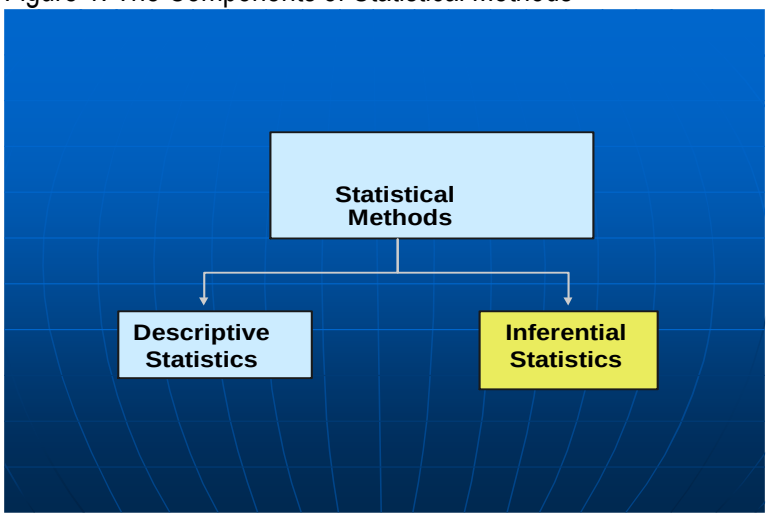
"Statistics as a discipline is the development and application of methods to collect, analyze and interpret data. Modern statistical methods involve the design and analysis of experiments and surveys, the quantification of biological, social and scientific phenomenon and the application of statistical principles to understand more about the world around us. Since data are used in most areas of human endeavor, the theory and methods of modern statistics have been applied to a wide variety of fields. Some disciplines that use modern statistical methods include the medical, biological and social sciences, economics, finance, marketing research, manufacturing and management, government, research institutes and many more" (http://statistics.unl.edu/whatis.shtml October 23 2007; the University of Nebraska-Lincoln Department of statistics.)

What Statistics is about: Statistics deals with drawing INFERENCE or conclusions about the entire POPULATION when only part of it is observed. In statistics, we deal with selecting representative SAMPLES, and deriving INFERENCES concerning POPULATIONS using samples.

Historically, statistics have been used to tell a story with numbers. Numbers usually communicate issues more succinctly than words do. The message that is carried using numerical data is often quite clear; for instance in 2004 Uganda's population was about 24 million. One does not need to look at many phrases while trying to get the current population estimate of Uganda. On the other hand Biostatistics is defined as a process of applying statistical tools to data derived from biological sciences.

Statistical Methods: Statistics is divided into two main components, ***descriptive*** and ***inferential*** statistics as shown in figure I.

Figure 1: The Components of Statistical Methods



The above diagram shows that statistical methods are grouped into two areas. These are aligned with the 'epidemiological approach'. Epidemiologists first observe and describe phenomena about disease or health related events. Thereafter, they examine the association between possible causal factors and disease. The same 2 step approach is applied in Biostatistics:

1. Descriptive Statistics: The aspects of organization, presentation and summarizing of data are labelled as Descriptive Statistics. This was the main subject matter of statistics up to the 20th century. In health sciences, descriptive statistics are used to characterize the health status of a particular population. Examples of descriptive statistics include measures of location or measures of central tendency and measures of dispersion that will be explained later.

2. Inferential Statistics: On the other hand, statistical inference is concerned with the logical basis by which conclusions regarding populations are drawn from results obtained from a SAMPLE. Estimation and hypothesis testing are examples of inferential statistics.

Activity 2.1.1
Give examples of how statistics has been used in the following area after a statistical exercise; (Finance, Health, Public policy)
List down all the statistical tools/components you may have come across earlier.

I am sure that the definition of statistics has helped you to identify basic statistical tools. Let us now consider the role of statistics in health sciences. Biostatistics is used for administrative decisions, planning, monitoring and evaluation; projections and answering research questions.

b. Some basic concepts

Like any other field of learning you will notice that statistics also has its own terminologies. Although some terms may appear to be familiar, they may have specialized meanings that are different from the meanings we usually associate with these terms. Some of the common terms used are:

- Data
- Variable

- Population
- Sample

1. Data: Data is a raw material of statistics. In statistical terms, we define data as a number. Also, we note that "data" comes from a latin word "*datum"* (something given) which is a past participle of *dare* (to give). Datum is an observation, value, given. Other text say data are observations or facts which when collected, organized and evaluated become information or knowledge. You will find out that in statistics, we use two kinds of numbers; those that result from counting i.e. number of patients being discharged from a hospital and another form are those numbers that result from measurements such as taking a patient's temperature or a nurse measuring a child's weight.

2. Variable: A variable is a characteristic that may assume more than one set of values to which a numerical measure can be assigned. For example, we can have demographic characteristics of individuals such as height, age, or any other characteristics of individuals such as the amount of income, province or country of birth, and grades obtained at school. Variables may be classified into various categories, such as;

Categorical (*Qualitative) variables:* These are variables that are intrinsically non-numeric. A categorical variable (also known as qualitative variable) is one for which each response can be put into a specific category. These categories must be either mutually exclusive or exhaustive. Mutually exclusive means that each possible survey response should belong to only one category (for example a respondent is said to either be "male" or "female" but not both!); exhaustive would require that the categories should cover the entire set of possibilities. Categorical variables can be either nominal or ordinal.

Nominal variables

A nominal variable is one that describes a name or category. This variable does NOT have 'natural ordering' of the set of possible names or categories. For example Sex and mode of transportation for travel to work (e.g car, bicycle, train …) are nominal because they describe the category of sex and transportation, respectively.

Ordinal variables

An ordinal variable is a categorical variable for which the possible categories can be placed in a specific order or in some 'natural' way. For example, the variable 'behaviour' is ordinal because the category 'Excellent' is better than the category 'Very good', etc. Also, ranks in the military (Captain, Major, Field Marshal) have order implied. However, the limitation may be in knowing by how much these categories differ. For example, we may not know by how much 'Excellent' behaviour is better than 'Very good' behaviour.

Numeric (or *Quantitative or Continuous)* variables

These are variables that are intrinsically numeric. A numeric variable, also known as a quantitative variable, is one that can assume a number of real values—such as age or number of people in a household. However, not all variables described by numbers are considered numeric. For example, when you are asked to assign a value from 1 to 5 to express your level of satisfaction, you use numbers, but the variable (satisfaction) is really an ordinal variable.

Random variable: A Random Variable is a variable associated with a random sample. If the process that generates the value of the variable is random, then the variable containing those values can be defined as a random variable. Alternatively, a random variable can be thought of as a function, which assigns unique numerical values to all possible outcomes of a random experiment under fixed

conditions; this definition implies that a random variable is not necessarily a variable but rather a function that maps events to numbers

Numeric variables may be either ***continuous random variables*** or ***discrete random variables***.

Continuous random variables

A variable is said to be continuous if it can assume an ***infinite number of real values.*** Examples of a continuous variable are distance, age and temperature. The measurement of a continuous variable is restricted by the methods used, or by the accuracy of the measuring instruments. For example, the height of a student is a continuous variable because a student may be 1.6321748755... metres tall. However, when the height of a person is measured, it is usually measured to the nearest centimetre. Thus, this student's height would be recorded as 1.63 m.

Note: To make them easier to handle, continuous variables are usually grouped into "class intervals", which is part of the process of organizing data so that they become useful information.

Discrete random variables

As opposed to a continuous variable, a discrete variable can only take a ***finite number of real values***. An example of a discrete variable would be the score given by a judge to a gymnast in competition: the range is 0 to 10 and the score is always given to one decimal (e.g., a score of 8.5) or counts of say number of daily travelers by car in a given town, number of students in MPH class, number of patients admitted in a cancer clinic. Discrete variables may also be grouped. Again, grouping variables makes them easier to handle.

Note: Measurement of a continuous variable is always a discrete approximation.
Information from Canada's National statistical agency,
http://www.statcan.ca/english/edu/power/ch8/variable.htm)

Note: A numerical variable can be categorized to create new qualitative variables. For example all people aged less than 15 years may be given category "0" and all those above 15 years can be categorized as "1", and such information put in a qualitative (categorical) variable called ***Agegroup.*** Although the variable "Agegroup" has numeric values, the variable is NOT intrinsically numeric, it is only holding numeric values identifying the two categories, which makes it a categorical variable.

Activity 2.1.2
From your own environment, outline 10 different variables

3. Population: A population is defined as a collection of all elements (or the complete list of ALL units of a group(s)) for which we have an interest at a particular time point. If we take a measurement of some variables on each entity in a population, we generate a population of that value or set of variables. Some examples of populations are the population of vehicles in Kampala, population of all medical school students and the population of children under five years who are malnourished. A ***Parameter*** is the unknown, quantitative measure (e.g., mean age, height, weight, income) for the entire population or for specified domains which are of interest to the investigator.

Note that the word population does not refer to human beings only but to anything of interest

4. Sample: This is a subset of the population, and can also be thought of as a collection of values of one or more variables of a given population. Suppose our population consists of all motorcyclists in Kawempe division. If we select a few of them and measure their sight, the selected cyclists are what we refer to as a sample. A random sample is that which is chosen such that each member of the target population has a chance of being selected. A **statistic** is a numerical characteristic (or quantitative measure) of a sample (e.g sample mean, median); usually we infer from a statistic to a parameter under certain assumptions.

c. Sources of data

In order to describe the status of particular population, you need to use statistics. Statistics are derived from data. You therefore need to begin by searching for suitable data to serve as your raw material for a particular investigation. Such data is usually available from one or more than one source. Different sources of data include:

- Routinely kept records
- Surveys (E.g Demographic and Health Surveys)
- Experiments
- External Sources
- Census (usual national censuses)

Details of these sources are discussed in the next sub sections. [These sources are further expounded on in the unit Demography and Population Dynamics]

1. Routinely kept records: Almost all institutions keep records on the day-to-day transactions of their activities. In health centres for example, both medical and financial records are kept. Medical records contain immense information on patients while financial records contain information on the facility's business activities. If you want data from a hospital for a particular reason or question, you should look first among the routinely kept records. Today all the health units in Uganda are supposed to summarize data on a monthly basis in their Health Management Information System forms (HMIS).

Activity 2.1.3 Visit any nearby health unit and find out the kind of information kept routinely
Activity 2.1.4 Contact the HMIS Focal Person in your District Health Office and obtain a copy of the Health Management Information System forms. Read through and acquaint yourself with the different types of reports and their scheduling.

From that activity, you should have noticed for example that all the necessary data on morbidity is either kept in that particular unit you have visited or summarized in the District HMIS register. From your visit you could have also discovered that reports are compiled on a monthly basis. Why do you think this is so? And who are the major consumers of this information?

2. Surveys: A survey can be defined as the collection of data by use of a standardized questionnaire that is administered by special interviewers or mailed to respondents. Surveys can be carried out on the whole population (census) or a sample.If data that required answering a question is not available from routinely kept records, another source one may use is a survey. Suppose you want to obtain information regarding patients' type of housing, income and distance from home up to the health unit, you may conduct a survey among patients to obtain this kind of information.

3. Experiments: A study where the researcher deliberately influences events and investigates the effects of the intervention. Examples of experimental studies include clinical and field trials. For instance if a doctor wants to find out about the best combination of malaria drugs, he has to design a number of experiments by looking at different anti-malarial drug combinations. Subsequent evaluation of the findings might enable the doctor to take a decision on the most effective drug combination. These study design can give stronger inferences compared to surveys, and are usually used to compare between groups.

4. External Sources: External sources of data are sometimes referred to as secondary data sources. As it sounds, external sources of data arise when data needed to answer a particular question already exists in the form of published reports or research literature. This implies that someone else has already gathered the information and the answers obtained may also be applicable to the present research question.

Activity 2.1.5: Identify and list the various external sources of data in public health.
Exercise 2.1.1 : Read Wayne W. Daniel pages 12 to 16 and do the exercise given on numbers 1 to 7

Lesson 2: Statistical Measurements and Scales

Introduction: In statistics, we measure or count variables and record the observations. Our measurements are expressed either as continuous attributes (e.g. weight 5kgs, 6kgs) or counts of categorizations (6 girls, 5 boys). In order to measure effectively, we make use of scales. The different types of scales will be expounded on in this lesson. We shall also recap our knowledge of rates, ratios and proportions [Refer to Unit 2: Applied Epidemiology and Biostatistics].

Lesson Topics: The following topics will be covered:-

1. Measurements and measurement scales
2. Ratios, Rates and Proportions

Lesson Objectives:

By the end of this lesson, the MPH STUDENT should be able to:

1. Contrast the different measurement scales used in statistics
2. Distinguish between rates, ratios and proportions and illustrate their application in summarizing data in biostatistics

a. Measurements and measurement scales

This section dwells a lot on the terms measurement and measuring scale. Measurement may be defined as the assignment of numbers to objects or events according to a set of rules based on value weighting. You will realize that various measurement scales result from the fact that measurements may be carried out under different sets of rules. Common scales of measurements include Nominal scale, Ordinal scale, Ratio scale and Interval scale. The nominal scale and the ordinal scale are described as *categorical scales;* the interval and ratio scales are described as *numerical scales.*

1. Nominal scale: This is the simplest scale of measurements. In this case numbers are used to represent certain categories that are mutually exclusive (by mutually exclusive we mean that if an item lies in one category, it cannot lie in another). The nominal scale deals with **categorical** data – For example, Gender (male-female), Marital Status (married-not married) and religion (catholic, Moslem etc…). Another characteristic of the nominal scale is that there is no implied order between categories i.e. there is no trend between categories. Numerical values of this scale are just labels,

Activity 2.1.5a: Identify 5 variables that can be measured using nominal scale.

2. Ordinal scale: The observations are grouped into ordered categories. They are not only different from category to category but can also be ranked according to some criterion. However, the differences in categories are not necessarily equal, and may not be meaningful, but order is implied. For example;

- Student evaluation (excellent, satisfactory, unsatisfactory)
- Frequency of attacks (rarely, sometimes, often times, very often)
- Disease status (moderate, mild or severe)
- Ranks if the army (Captain, Major , General etc…)

Activity 2.1.5b: Identify 5 variables that can be measured using ordinal scale.

3. Interval scale: In this case the Interval or distance between two points has precise meaning, but the zero point is arbitrary. Addition or subtraction is allowed in the interval scale. Let's consider a situation

where temperature is measured in either Fahrenheit or Celsius, the unit of measurement is the degree of heat and the point of comparison is the arbitrary chosen "zero degrees" which does not indicate lack of heat. This kind of scale is truly quantitative unlike ordinal and nominal scales of measurements discussed above.

4. Ratio scale: As you will explore, this the highest scale of measurement. Under this scale, equality of distances or ratios can be established easily. This scale has a true zero point. Intervals or distance between two points has precise meaning with a true or meaningful zero point. In this scale we have the ability to compare directly because there is a true zero. If you are 1.6m tall and your child is 0.8 m, then you are twice as tall. Other examples of measurements that make use of the ratio scale include; weight, length and height. It is the most commonly used scale of measurement.

b. Rates, ratios and proportions

As you go further in the field of biostatistics, you will realize that these three measures (rates, ratios, and proportions) are commonly used almost all the time to summarise data obtained from measurement of categorical variables. Before you learn about specific measures, it is important to understand the relationship between these three types of measures and how they differ from each other. All three measures are based on the same formula:

Ratio, Proportion, Rate = $x/y * 10^n$

In this formula, x and y are the two quantities that are being compared. The formula shows that x is divided by y . 10^n is a constant that we use to transform the result of the division into a uniform quantity. 10^n is read as "10 to the nth power". The size of 10^n may equal 1, 10, 100, and 1,000 and so on depending upon the value of n. You will learn what value of 10^n to use when you learn about specific ratios, proportions, and rates [in the Courses: Applied Epidemiology I and Demography and Population Dynamics]

1. Ratio: In a ratio, the values of x and y may be completely independent, or x may be included in y . For example, the sex of university students attending VCT services in a University hospital could be compared either as *male/female* or *male/ (male + female)*. In the first option, x (female) is completely independent of y (male). In the second, x (female) is included in y (all). Both examples are ratios.

2. Proportion: This is the second type of frequency measure used with dichotomous variables, in which x is included in y . Of the two ratios shown above, the first is not a proportion, because x is not a part of y. The second is a proportion, because x is part of y . {male/ (male + female)}

3. Rate: Rate is the third type of frequency measure used with dichotomous variables and it is often a proportion, with an added dimension: it measures the occurrence of an event in a population over time. The basic formula for a rate is as follows:

Rate =

$$\frac{\text{Number of cases or events during a given time}}{\text{Population at risk during the same period}} \times 10^n$$

Note that there are three important aspects of this formula.

1. The persons in the denominator must reflect the population from which the cases in the numerator arose

2. The counts in the numerator and denominator should cover the same time period
3. In theory, the persons in the denominator must be "at risk" for the event, that is, it should have been possible for them to experience the event.

2.1.7 Extension Activities

Extension Activity 1: Discussion Forum Question
NO DISCUSSION QUESTION

Extension Activity 2: Self- Assessment Quiz

QUIZ 2.1.1

(Choose the most correct option)

1. The following statements summarise what statistics is about, except one. Indicate the incorrect statement

a. Collection, collation, presentation and analysis of data
b. Drawing inference about entire populations when only a part is studied
c. Selecting representative samples and deriving representative parameters from them
d. Attributes that take on different values in different persons, places or time

2. Characteristics that cannot be counted, but can be categorised are often referred to as:

a. Quantitative variables
b. Discrete random variables
c. Qualitative variables
d. Discrete variables

3. One of these statements is true about statistical measurement scales

a. In the nominal scale, observations are chunked into categories with hierarchical order
b. In the interval scale, there is a true zero point and the interval between two points has precise meaning
c. In the ordinal scale, there is no implied order between categorisations and categories are not necessarily equal
d. In the nominal scale, categorisations are mutually exclusive and independent

4. X, Y and Z are dichotomous variables in which X is included in Y but Y is mutually exclusive with Z (an item in Y cannot be included in Z). If we express X as a percentage of the sum of Y and Z, at a precise point in time, this is then an example of:

a. A Proportion
b. A Ratio
c. A Rate
d. A Fraction

2.2 Unit 2: DESCRIPTIVE STATISTICS

2.2.1 Introduction to the Unit

"Descriptive statistics are used to describe the basic features of the data in a study, and to summarize the attributes of the population under study. They provide simple summaries about the sample and the measures", and can include simple tabulations and graphics" (http://www.socialresearchmethods.net/kb/statdesc.php). Therefore, descriptive statistics simply describe what is or what the data shows. Thus summarizing data, and meaningful organization and presentation to describe data are labelled as descriptive Statistics. In the health sciences, descriptive statistics are used to characterize the health status of a particular population, or the 'burden of disease or health related events'. Examples of descriptive statistics include measures of location also know as measure of central tendency, and measures of dispersion or spread or variation that will be explained in this unit. Very often, epidemiological inquiry starts with observation and description of phenomena. This is the basis for deciding whether a particular phenomenon is occurring beyond what is expected in time place and person. For many studies, this may be adequate to provide the information needed for interventions to address the anomaly. To describe phenomena in populations, we use descriptive statistics. These will be the subject of this unit.

2.2.2 Unit Outline

The following topics will be covered:

1. *Measures of location*
2. *Measures of dispersion*
3. *Application of measures of location; discussion and limitations*
4. *Counts and proportions*
5. *Other descriptive measures*
6. *Tabular and graphical presentations*

2.2.3 Instructional goal

The student should develop the competency to employ descriptive parameters in summarizing and presenting statistical data to those who need it.

2.2.4 Unit Objectives

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

1. Calculate and interpret measures of location/central tendency.
2. Choose and apply appropriate measures of location to summarize data.
3. Calculate and interpret the measures of dispersion.
4. Choose appropriate graphs, plots and tables and employ them to display data in a way that simplifies interpretation and use.
5. Select appropriate statistic that can be used to describe and summarize categorical data.

2.2.5 Time Frame

1 WEEK

2.2.6 Content

Lesson 1: Measures of Location/Central tendency and Dispersion/spread

Introduction: In real life, every time you meet a person, the characteristics that describe him/her beam at the back of your head. If you don't have any information about the person, you may want to know the name, origin, age and other characteristics depending on the circumstances and need. The same thing applies to data. Take an example of a data set with birth weight from a nutritional survey. You need single summary measures that can describe the weight of the population or sample.

There are two kinds of measures used; one identifies the centre of the data or a value that represents most of the sample/population units (Measures of location) and another identifies how the data is distributed or how far each value is from a measure of location (measure of dispersion). The commonest measures of location are the mean, mode, median while measures of dispersion include the range, standard deviation (SD), variance and coefficient of variation (CV). Each measure has its own place in summarizing public health data. These measures apply to continuous data when such data are ungrouped (e.g. height, age). However, continuous data may be **grouped** (e.g. age ranges). The purpose of the lesson is to make you have a feel of what goes on in the computers or calculators and interpret the results.

Lesson Topics: The following topics will be covered:

1. Summarising numerical data
    a. Measures of location: The mean, the mode, the median, the geometric mean
    b. Measures of dispersion: The standard deviation, the variance, the coefficient of variation, ranges.
2. Application of measures of location/dispersion and their limitations
3. Summarizing categorical data: Counts, frequencies and proportions

Lesson Objectives:

By the end of this lesson, the student should be able to:

1. Calculate and interpret the following measures of location: -mean, mode, median, geometric mean and quartiles
2. Choose and apply appropriate measures of location to summarize data
3. Calculate and interpret the following measures of dispersion: Standard deviation, Range, Variance

a. Summarizing Numerical Data

When you have a set of numerical data from a population, how do you summarise it? You summarize numerical data using 2 types of measures – measures of location and measures of dispersion.

A1. Measures of Central tendency/location

The common measures are the Mean (arithmetic or geometric mean), the Mode and the Median. These summarize the data into a single measure that provides a picture about the sample or population.

i) **The arithmetic Mean:** This is also called the simple mean, an average or expected value. It is used to summarize intrinsically continuous data whose scale of measurement is either interval (where equal distances between values, but zero point is arbitrary, e.g. Temperature and IQ) or ratio (where equal intervals between values and a meaningful zero point, e.g. height, weight, BP). It is usually represented

as $\bar{x}$. How is it determined?: For ungrouped data, it is the sum of all observations ($\sum_{i=1}^{n} x_i$) divided by the number of the observations in a sample (n).
The formula for calculating the mean is therefore written as:

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

An example is systolic blood pressures (mmHg) for five (5) individuals visiting a health facility.

Five systolic blood pressures (mmHg) (n = 5)
− 120, 80, 90, 110, 95

$$\bar{x} = \frac{120 + 80 + 90 + 110 + 95}{5} = 99mmHg$$

The sum of the systolic pressure ($\sum x_i$) =120+80+90+110+95=495

The mean age $\bar{x}$ is $495/5$= 99mmHg

Note: when dealing with the whole population the number of observations, N, is the total of all samples whose sizes are n_i; the formula for the total number of observations in the population is a summation of the samples of sizes n_i; N= $\sum_{i=1}^{\infty} n$

Disadvantage of the mean: The mean is the most commonly used measure of location. It has one main limitation: It is sensitive to extreme values; i.e. values that are usually much smaller or much larger than most of the data. An example would be additional systolic pressure values of say 140 will now make six observations in the above example, and the new sum of the dataset now is ($\sum x_i$) =120+80+90+110+95+140=635

The new mean $\bar{x}$ is $\frac{635}{6}$=105.8 mmHg. You can see the mean has increased by 6.8! The mean is therefore not a good estimate when the data has outlier observations or generally when data are skewed in either one of the directions (left or right).

Weighted average: Sometimes a dataset contains a number of repeated values or we may not have all the observations but instead have a frequency table. We can then estimate the mean by a weighted average: by multiplying each data value with the number of observations that have such a value, adding the products and dividing by the total number of observations (total of the frequencies).
The formula for computing the weighted average is:

$$\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Where; w_i is the weight associated with observation x_i, the frequency with which it appears.

This is demonstrated below:

Observation, *x*	Frequency, *f*	Product, *xf*
20	3	60
26	1	26
28	1	28
32	1	32
40	2	80
$\sum$	**8**	**226**

The mean in this case is estimated by: $\frac{226}{8} = 28.5$

Exercise 2.2.1: Compute the mean of the heights (metres) of women in ANC clinic- 1.6, 1.5, 1.4, 1.6, 1.5, 1.7, 1.55

ii) The Median: This means "middle". It is the middle of a set of observations when arranged in either ascending or descending order. It has no conventional symbol (sometimes expressed as M or md)

In our example, the systolic blood pressure 120,80, 90, 110, 95 be arranged as
80, 90, 95,110, 120; here, you can see that the middle value is **95 mmHg**.
a) Odd number of observations: In some cases the number of observations will be odd
If the number of observations is odd, the median is the middle observation after arranging in either ascending/descending order
b) Even number of observations:
If the number of observations is even, then the average of the two middle observations will constitute the median. E.g. 120,80, 90, 110, 95, 140
Arrange in ascending order:
80, 90, *95,110*, 120, 140
The middle values are indicated in the brackets 80, 90, **(95,110),** 120, 140
Find the average of these middle two values: **(95+110)/2=102.5 mmHg as the median**

For ungrouped data, the median can therefore be identified using the following steps:

- Arrange the series in ascending or descending order
- Find the middle rank of the observations using the formula:

  Mid rank = $\frac{(n+1)}{2}$, n=number of observations
- If n is odd, the middle rank falls on an observation. If n is even the middle rank falls between two observations
- Identify the value of the median
- If the middle rank falls on a specific observation then the median is equal to the value of that observation.
- If the middle rank falls between observations (that is if n is even) the median is the average of the values of the observations at $\frac{n}{2}$ and $\left(\frac{n}{2}\right)+1$.

In the example of systolic blood pressure: 80, 90, 95,110, 120, 140 the middle rank is (6+1)/2=3.5. The value of median is therefore between the 3rd and 4th observations. This will be (95+110)/2=102.5 The median is therefore 102.5mm Hg.

Characteristics: An advantage with the median over the mean is that it is not easily influenced by extreme values. In the example of 8 children the mean and median are 28.3 and 27 respectively. On addition of one child with 60 months, the mean and median become 31.5 and 28 respectively. The mean increased by 3.5 while the median increased by only 1.

Activity 2.2.1: Determine the median height in metres of 7 women attending an ANC clinic 1.6, 1.5, 1.4, 1.6, 1.5, 1.7, 1.55
Activity 2.2.2: Determine the median height in metres of 8 women attending ANC clinic 1.72, 1.6, 1.5, 1.4, 1.6, 1.5, 1.7, 1.55

iii) The Mode: For ungrouped data, this refers to the commonest value in a list of observations. In the example of Systolic in mmHg for the 7 participants :

If more persons had Systolic blood pressure taken, n=7 persons: 140, 90, 80,95,110, 90, 120
The mode is 90 because it occurs most often. It occurs 2 times, which is the highest number of repetitions. Creating a frequency distribution table more easily identifies the mode.

In the table below it is easier to see 90 as the mode since it occurs more often than any other systolic pressure value.

Observation	Frequency
80	1
90	2
95	1
110	1
120	1
140	1

Exercise 2.2.2: Determine the mode of the following heights of women attending ANC clinic
1.72, 1.6, 1.5, 1.4, 1.6, 1.5, 1.7, 1.55

Additional Characteristics of the mode: A set of values may have more than one mode, when the set has two modes, it is called *bi-modal.* The mode can also be observed in a frequency table for grouped data. In this case, it equates to the modal class.

iv) The Geometric mean:
Non-zero and non-negative data tend to be positively skewed. E.g. Laboratory data such as iron levels in blood, Viral load, CD4. Such data therefore do not usually have a normal distribution; so arithmetic mean is not a good summary measure to use. Instead a summary measure known as geometric mean is used. The geometric mean is the anti-logarithm of the arithmetic mean of a set of data measured on a logarithmic scale. The geometric mean can also be defined as the n^{th} root of a product of n non-zero non-negative observations.

The arithmetic mean does not clearly describe data that has a skewed distribution. The geometric mean can instead be computed. The raw data is thus mathematically transformed to have a more symmetric distribution; with the logarithmic transformation being the most frequently used. If raw data are transformed into a logarithmic scale and the arithmetic mean computed, the arithmetic mean of the

transformed data will still be in a logarithmic scale. However, if we anti-log (or transform the arithmetic mean of the logarithmic data back to the raw data scale), we obtain a geometric mean, as in the first definition.

Note: Geometric mean is NOT the same as the arithmetic mean! Actually the geometric mean is used on data that are skewed, with non-zero and non-negative (positively skewed) values so that we can obtain the logarithms of the observations.
Using the second definition, we can calculate the geometric mean from individual measurements in a dataset is:

$$\bar{x}_{geo} = \sqrt[n]{x_1 \times x_2 \times .... \times x_n}$$

Using the first definition of log-transformation of the raw data, we obtain the geometric mean as:

$\bar{x}_{geo}$ =antilog $\left(\frac{1}{n}\sum Logx_i\right)$,

where $\left(\frac{1}{n}\sum Logx_i\right)$ is the arithmetic mean of the log-transformed data.

It is this arithmetic mean of the log-transformed data that will be anti-logged back to the scale of the raw data, and then called the geometric mean.

In other words you change the observations (x_i) into a logarithm base 10 or natural log. It doesn't matter which logarithm you choose. Then add the logs and divide by the number of observations to get the mean of the log-transformed data. The geometric mean is the antilog of the value obtained.
Examples of data where geometric mean is useful Viral loads, CD4 levels because they tend to have skewed data with non-zero non-negative values.

Example of a Geometric mean calculation
A blood draw for CD4 assessment from four HIV+ men seeking care at a health facility in Kampala are provided below
CD4 cell data, X= 2 400 635 1302
Step 1
Log transformation of the X values

Patient_ID	CD4, X	y =$Log_{10}X$
1	2	0.301
2	400	2.602
3	635	2.803
4	1302	3.115
Arithmetic Mean	584.8	2.205

Then find the antilog(y)=antilog(2.205)=160.4

The Geometric mean (x_{geo}) is the antilog (y)=160.4

Example 2.2.1: Number of people that have received ARV drugs over two year intervals in a country are:

10, 100, 100, 10,000, 10,000, 100,000 1,000,000

1. The first step is to convert them to log. In this case we can take base 10.
Remember

$10=10^1$
$100=10^2$
$1000=10^3$
$10,000=10^4$
$100,000=10^5$
$1,000,000=10^6$

The $\log_{10}(x_i)$ =1, 2, 2, 4,4,5,6

2. Calculate the mean of the logs: Mean of $\log_{10}(x_i)$ = (1+2+2+4+4+5+6) =3.43

The anti-$\log_{10}(3.43)$ =2511.9; the geometric mean number of people that have received ARV is 2512.

Exercise 2.2.3: Calculate a geometric mean for the number of patients reporting obesity related complications in a hospital: 2, 2, 4, 8, 8, 16, 16, 16, 32, 64

vi) Mean for grouped data

Useful summarisation can be obtained by grouping data. For grouped data, the mean is obtained using the formula:

$$\overline{x} = \frac{\sum_{i=1}^{n} fx_i}{\sum_{i=1}^{n} f_i}$$

Where: f is the frequency in each category and x_i is the mid-point of the category.
It is demonstrated as follows:

Score (Category)	Freq (f)	Mid-point of interval (x_i)	fx_i
41-50	1	45.5	45.5
51-60	3	55.5	166.5
61-70	8	65.5	524
71-80	3	75.5	226.5
81-90	2	85.5	171
91-100	3	95.5	283.5
Total	20		1420

Mean for the grouped data = $\overline{x} = \frac{1420}{20} = 71$

vii) Median for grouped data

The formula to obtain the median for grouped data is given by:

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

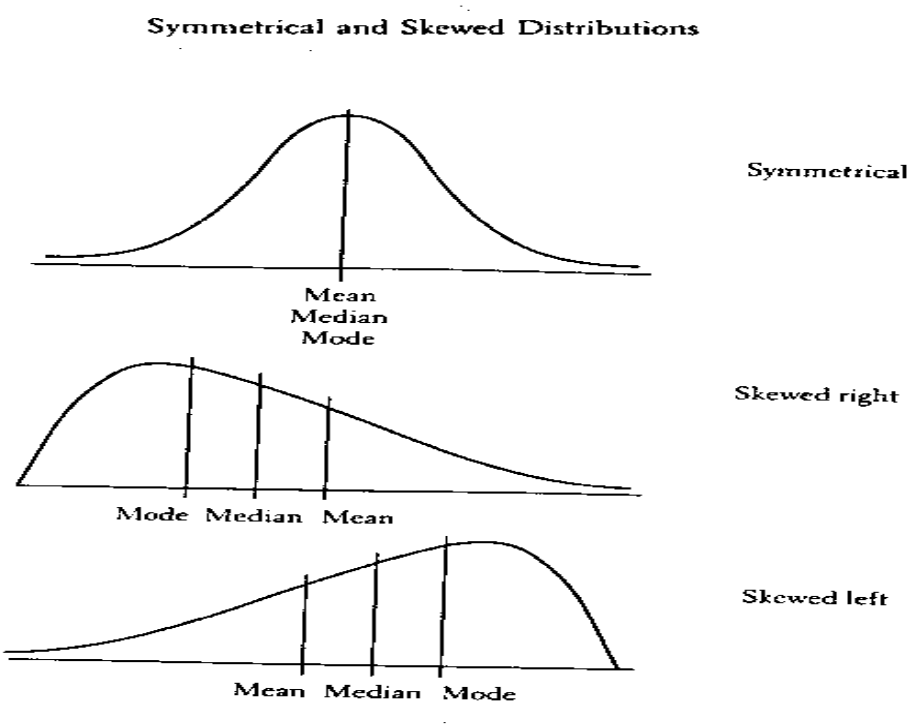
Where: L= lower limit of the class containing the middle case

N=total population
F=cumulative frequency up to age group before the one containing the miffle cae
f=frequency of the class containing the middle case
i=size of the class interval contining the middle case

viii) Relationship between the mean, mode and median in symmetrical and asymmetrical distributions

The value of the mean, the mode and the median can inform us of the degree of symmetry in a data series. When the mean, the median or mode are equal or nearly equal, we have symmetrical or near symmetrical distribution. When the median is less than the mean the data set is skewed to the right (the tail is on the right); when it is more than the mean the data set is skewed to the left (the tail is on the left) [see figure 3.1 below]. As we shall see later, symmetry of data is very important in estimation and hypothesis testing. Most parametric tests that we shall see later on in the course are under lied by an assumption that the data in symmetrical.

Figure 4.1: Symmetry and Skewed distributions



Source: *http://business.clayton.edu/arjomand/business/l3.html*

v) Summary: Measures of location are single values that summarize quantitative observations. The most common measures are mean, median, quartiles and mode. The mean is ideal when data is or nearly normally distributed. The median is better with skewed data because it is not very sensitive to extreme values. When data follows a logarithmic or exponential pattern then geometric mean is the most suited summary measure.

A2. Measures of Dispersion

A graph showing distribution of data normally shows a peak near the centre and a spread outwards. Just as we use measures of location to identify the centre of the data, measures of dispersion show the extent of the spread outwards. They are also called measures of variation or variability. If data is widely scattered, the dispersion is greater. The commonest measures of dispersion are the range, the standard deviation and the coefficient of variation (CV).

i) Range: The range is the difference between the highest (Maximum) and lowest (Minimum) value in a set of observations. In epidemiology the range is often reported by the minimum and maximum values such as "from (minimum) to (Maximum)" in the example of height in metres of women: **1.72, 1.6, 1.5, 1.4, 1.6, 1.5, 1.7, 1.55**, we demonstrate how to find the range;

Step 1: Arrange data in ascending/descending order

1.4, 1.5, 1.5, 1.55, 1.6, 1.6, 1.7, 1.72

Step 2: Identify the minimum and maximum

Minimum =1.4 Maximum=1.72

Step 3: Calculate the range – difference between the minimum and maximum figure.

1.72-1.4=0.32

Exercise 2.2.4: Compute the range of the following observations of age in months of kids in a malaria clinic: 29, 20, 40, 32, 26, 28, 20, 20, 40

ii) Percentiles, quartiles and Inter quartile range: A value is said to be an n^{th} percentile in a given set of observations if n percent of the observations fall at or below it. In other words a percentile is any of the 99 values that divide the values in a set of data into 100 equal parts, so that each part represents 1/100th of the sample or population. Common percentiles are the 25th, the 50th and the 75th. A 50th percentile is a median. The 25th and 75th percentiles are also known as lower and upper quartiles respectively. The 50th percentile (median) is a measure of location while other percentiles are normally used to gauge the spread of the data.
One of the commonest use of percentiles is in computation of the inter quartile range. It is the difference between the 25th and 75th percentiles. It describes the middle 50% of the observations. As a measure of spread, the interquartile range is not influenced by extreme values as is the range.

Example 2.2.2: The distance in kilometres of 11 health units from the district headquarters is 6, 47, 49, 15, 43, 41, 7, 39, 43, 41, and 36. We can calculate the inter quartile range from lower and upper quartiles using the steps below

Step 1: Arrange in ascending order

6, 7, 15, 36, 39, 41, 41, 43, 43, 47, 49

Step 2: Identify the position of the lower and upper quartile

Position of lower quartile $Q_1 = \left(\frac{n+1}{4}\right) = 3$

Position of upper quartile $Q_2=\left(\frac{3(n+1)}{4}\right)=9$

Step 3: Identify the values that occupy positions Q_1 and Q_2.
Like in a median, if a value occupies the position Q_1 then the value is the lower quartile. Similarly, if it occupies position Q_2 then it is the upper quartile.

The value for position Q_1 =15 and the value for position Q_2=43.

Step 4: Calculate the inter-quartile range
Subtract the value on position Q_1 from the value on position Q_2: 43-15=28.

If the quartile lies between observations, the value of the quartile is *the value of the lower observation plus the specified fraction of the difference between the observations*. For example if the position of a quartile 20¼ it lies between 20th and 21st observations. So the quartile is the value of the 20th observation plus ¼ the difference between the 20th and 21st observations.
Suppose we had another distance of a health unit as 51km.

Then Q_1=3¼ and Q_2=9.75

- Then the lower quartile would be in position between the 3rd (15) and 4th (36) values; this will be 15 + ¼ X (36-15) = 20.25
- The upper quartile will be between the 9th (43) and 10th (47) values; this will be 43+ ¾ X (47-43) =46.
- The inter quartile range = 46-20.25 = 25.75

Approach two
The statistical package STATA uses this approach.
Given a set of observations, n, we can easily find the median as explained above. However, for the lower (25% of the data) and upper (75% of the data) quartiles we follow these steps below;

Inter-quartile range (IQR)
Suppose a total of five young boys took height and provided below:

- If height (meters)= 1.50, 1.40, 1.56, 1.64, 1.60
  Because the number of males, n, is odd then

If n=odd

Step 1: Determine the median

Arrange the data in order (1.40, 1.50, 1.56, 1.60, 1.64)
The middle value of the arranged data/median is 1.56

Step 2
Split the data into two halves, but make sure the median falls in each half of the data
The lower half of the data is 1.40, 1.50, 1.56
The median of the lower half is 1.50, which is the Q1 (the median of the lower half of the data)

Step 3

The upper half of the data is 1.56, 1.60, 1.64
The median of the upper half is 1.60, which is the Q3 (the median of the upper half of the data)

The Inter-quartile-Range (IQR) is

- IQR=Q3-Q1=1.60-1.50=0.10

- Inter-quartile range (IQR)
- If $height_i$ (meters)= 1.40, 1.50, 1.56, 1.60, 1.64, 1.78

If n=even
Step 1: Determine the median
Arrange the data in order (1.40, 1.50, 1.56, 1.60, 1.64, 1.78)
The middle value of the arranged data/median is: The average of the 2 middle values form the median, (1.56+1.60)/2=1.58

Step 2
Split the data into two halves
The lower half of the data is 1.40, 1.50, 1.56
The median of the lower half is 1.50, which is the Q1 (the median of the lower half of the data)

Step 3
Split the data into two halves
The lower half of the data is 1.60, 1.64,1.78
The median of the upper half is 1.64, which is the Q3 (the median of the upper half of the data)

The inter-quartile-Range (IQR) is
Q1=1.50
Q3=1.64

QR=Q3-Q1=1.64-1.50=0.14

Activity 2.2.3: Find the inter-quartile range for heights (in metres) of women attending ANC clinic: 1.72, 1.6, 1.5, 1.4, 1.6, 1.5, 1.7, 1.55

iii) Variance and Standard Deviation: Variance and standard deviation measure how far observations are from the expected value or the mean. Dispersion is less when observations lie close to their mean. The variance is obtained by summing up squared differences of the observations from the mean and dividing by $n-1$, where n is the number of observations. The differences add to zero and when they are squared they all become positive numbers. The variance represents squared units in order to get back to the original units, we take the positive square root which is the **standard deviation (SD)**. The standard deviation is the most commonly used measure of statistical dispersion. It is non-negative and has the same units as the data. The standard deviation of the population is symbolized by σ while for the sample it is designated as 's'.

The steps for calculating the **variance** are:
1. Calculate the mean $(\bar{x})$
2. Compute the difference between each observation from the mean $(x_i - \bar{x})$

3. Square the differences – when we square the differences, we eliminate the negative signs and therefore, our sum cannot be zero $(x_i - \bar{x})^2$
4. Get the sum of the squared differences $\sum(x_i - \bar{x})^2$
5. Since the data is a sample, divide the sum (from step 4) by the number of observations minus one, i.e. $(n-1)$ (where n is equal to the number of observations in the data set). [The term $(n-1)$ will later be called **degrees of freedom**]. When we do so, we obtain the sample **variance, usually given as s^2**. On the other hand the population variance, σ^2 is obtained by dividing the sum of the squared differences by the total number of observations, or population size, N
6. Since we have all along been dealing with squared differences, we now obtain the square root of the variance. The **standard deviation** is the square root of the **variance**.

Remember! The population **variance** is the mean of the squared differences of all values from the mean of the observations. As you can see, if the differences were not squared, their sum would be zero. It gives us a picture of how far the values are distributed away from their mean i.e. a measure of the extent of dispersion.

Illustration: Calculation of the standard deviation and variance

Clinic	***Out patients In 9 clinics***	***Deviation from mean***	***Square Deviation***	***Computation***
	(x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	
Kawempe	20	-29	828	n =9
Nagulu	30	-19	353	$(n-1)$ =8
Komamboga	32	-17	281	
Nakawa	40	-9	77	Sum of squares= $\Sigma(x-\bar{x})^2$ =3316
Rubaga	44	-5	23	
Makindye	60	11	126	
Makerere	63	14	202	$\frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{3316}{8}$
Bwaise	70	21	450	Variance,S^2 = =414.4
Kalerwe	80	31	975	Standard Deviation = $S = \sqrt{Variance}$
Sum (Σx)	439	0	3316	$S = \sqrt{414.4}$ =20.4
Mean $(\bar{x})$	48.8			

Source: *Not real.*

Steps

1. Calculate the mean

   Σx $(\bar{x})$ = Σx /n=439/9=48.8

2. Compute the difference between each observation from the mean $(x - \bar{x})$
3. Square the differences

   $(x - \bar{x})^2$

4. Get the sum of the squared differences

   $\Sigma(x - \bar{x})^2$ =3316

5. Since the data is a sample divide the sum (from step 4) by the number of observations minus one,

i.e. $(n-1)$ Variance, $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{3316}{8}$ = 414.4

6. Standard deviation, S = $\sqrt{Variance}$ = $\sqrt{414.4}$ = 20.36

Just imagine if we were to rely on $(x_i - \bar{x})$ as the variation of outpatient attendance from the mean! We would end up with zero after computing the average. We say the standard deviation of the data set is 21 outpatients.

The Standard Deviation in Large data sets:
With a large data set, it is very cumbersome to calculate all the differences and square them; there are formulae we can use to compute the SD in a faster way. They give exactly the same results as above but are less time consuming. The computation we have gone through is meant to help you understand the standard deviation as a measure of variation. The simplified formulae are:

$Variance = S^2 = \frac{n\Sigma x^2 - (\Sigma x)^2}{n(n-1)}$; And the standard deviation, $S = \sqrt{\frac{n\Sigma x^2 - (\Sigma x)^2}{n(n-1)}}$

i.e. the following steps can be used:

1. We square all the individual observations and sum them up to obtain the sum of squares for all observations: $(n\sum x^2)$
2. We also obtain the sum of all observations and square it: $(n\sum x)^2$
3. We subtract the squared sum of observations from the sum of squares for all observations: $(n\sum x^2) - (n\sum x)^2$
4. We divide this by $(n-1)$ to obtain the variance
5. The SD is the square root of Variance

Illustration: Calculation of standard deviation

Clinic	**Out patients** In 9 clinics (x_i)	**Square of** (x_i) $(x_i)^2$	**Computation**
Kawempe	20	400	n=9 n-1 =8
Nagulu	30	900	Sum of x_i: $\sum x_i$ =439
Komamboga	32	1024	
Nakawa	40	1600	Sum of x_i^2: $\sum x_i^2$ 24729
Rubaga	44	1936	Variance
Makindye	60	3600	$\frac{n\Sigma x^2 - (\Sigma x)^2}{n(n-1)}$ = $\frac{9 \times 24729 - (439)^2}{9(8)}$ =414.4
Makerere	63	3969	
Bwaise	70	4900	
Kalerwe	80	6400	Standard Deviation $S = \sqrt{414.4}$ =20.4
Sum	439	24729	
Mean $(\bar{x})$	49		

Source: *Not real*

Calculate the mean: This is equal to 49

Calculate the sum of all observations $\sum x_i$ =439

Compute the square of each observation: x_i^2

Compute the sum of the square of the differences: $\sum x_i^2$

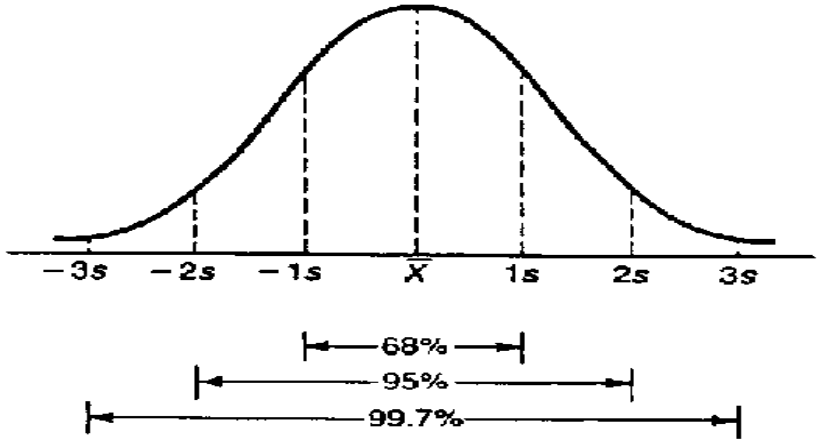
Square the sum of the observations: $(\sum x)^2$

Compute the variance

$$= \frac{n\Sigma x^2 - (\Sigma x)^2}{n(n-1)} = \frac{9 \times 24729 - (439)^2}{9(8)} = 414.4$$

Compute the standard deviation $S = \sqrt{414.4} = 20.4$

The Standard Normal Distribution and SD: As a rule, when data is normally distributed, approximately 68% of the observations will lie within one standard deviation, 95% will lie within 1.96 (~2) standard deviations and 97.7% lie within three standard deviations. See the figure 3.2 below



In the example above, if our data was normally distributed, in 68.8% of our clinics the number of outpatients would be within a range of 20 above or below the mean (49). That is from 29 to 69. In our example, 6 out of 9 facilities (66.7%) are within 1 standard deviation (20) from the mean.

Activity 2.2.4: Compute the variance and standard deviation for age in months of children in a malaria clinic 29, 20, 40, 32, 26, 28, 20, 20, 40

v) The Coefficient of variation: This is a relative measure of spread which expresses the standard deviation as a percentage of the mean. The standard deviations of two variables measured on different scales cannot be compared to each other in a meaningful way to determine which variable has greater dispersion e.g. comparing Blood Pressure and the Pulse Pressure (difference between systolic and diastolic BP), also known as the *vasalva ratio*. The standard deviations of these two cannot be compared because they are on different scales. However, since the standard deviation and mean are measured in the same unit and scale, and once divide the units cancel out thus producing a unitless measure. The coefficient of variation accounts for such differences between variables being compared in order to have a meaningful comparison. The COV is given by:

Standard Deviation (SD) X 100

Mean $(\bar{x})$

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/coefficient_of_variation.htm

iv) Conclusion: Measures of dispersion quantify the spread or variability of the observed values of a continuous variable. The simplest measure of the dispersion is the range from the smallest value to the largest value is a data array. The range is quite sensitive to extreme values. The Inter quartile range is preferred when the data is skewed.

For data, which is normally distributed, the mean is used together with the standard deviation. The standard deviation shows how observed values deviate from the mean. Sixty eight percent of the observations lie within one standard deviation while 95% of the observations lie within two standard deviations. For data, which is not normally distributed, the median and inter-quartile range is used to show the variation from the central point.

b. Summarizing Categorical Data

When you have a set of categorical data from a population, how do you summarise it? You summarize categorical data using counts (or frequencies) and proportions. For example, data from counts and proportions can be summarised as follows:

Variable	Variable	Number (or frequency)	Proportion (e.g. percentage)
Sex	Male	44	50.0%
	Female	44	40.0%
Religion	Christian	70	79.5%
	Moslem	10	11.4%
	Other	8	9.1%

Lesson 2: Presentation of Descriptive Data

Introduction: Suppose that we have 12 students in a bio-statistics course who have each achieved a score in a knowledge test. We can present this information in a straight forward and simplistic way e.g. arranged in order from the smallest to the highest as: **61, 69, 72, 76, 78, 83, 85, 85, 86, 88, 93 and 97**. However, this poses problems as it is:

- Too detailed
- Too broad
- Difficult to interpret
- Plausible for small data sets

There are available methods for summarizing this information in a way that the data set can be presented more meaningfully. We shall discuss these methods in this lesson.

Lesson Topics: The following topics will be covered:-

1. Other descriptive measures
2. Tabular and graphical presentations

Lesson Objectives:

By the end of this lesson, the STUDENT should be able to:

1. Choose appropriate graphs and tables and employ them to display data
2. Examine and interpret data that has been presented in the form of tables and graphs and draw relevant conclusions
3. Select appropriate parameters to describe categorical data

a. Other descriptive measures

In the previous lesson, we mainly looked at summarizing information that is continuous in nature, either grouped or ungrouped. We can also summarize nominal and ordinal data (categorical data) using numbers. We use:

1. Frequencies
2. Ratios, rates and proportions [These have been described in Unit 1; lesson 2]

These measures can also be used for numerical data that has been grouped e.g. age categories. A commonly used form of proportions is percentages.

b. Tables and graphical presentations

We can use tables and graphs to display data. This depends on whether it is numerical or categorical. We can employ:

1. Frequency tables: We subdivide numerical data into classes e.g. age groups, and indicate the counts in each group.

2. Histograms: They mainly show area. The continuous variable of interest is on the x-axis, usually in grouped form based on ranges. The size of the ranges must be uniform. The frequency of occurrence is on the y-axis.

3. Frequency polygons: A derivative of histograms in which a line is drawn to indicate the frequencies. They are preceded by a frequency table. They are useful when comparing two distributions on the same graph.

4. Line graphs: They indicate the variation of one continuous variable with another. They are usually used to display information that changes over time.

5. Scatter plots: They are similar to line graphs, and indicate the variation of a continuous variable with another. However, the plots are scattered about the line of best fit due to random error.

6. Stem and Leaf plots: In such plots, data are presented in form of a "stem" and "Leaf". The summarizing digits of the display constitute the stem, while the more varying digits represent the leaf.

7. Box and whisker plots: Data are divided into a box and whiskers. It is useful for comparing different sets of sampled data, to gauge their spread about a population mean. To construct a box and whisker plot, we do the following:

- We arrange the data set from the smallest observation to the largest
- We draw a uniform box over the inter-quartile range
- We draw whiskers outward from the box, to cover the parts of the data that are outside the inter-quartile range

Activity 2.3.5: Please read up and find examples of each of the following (Refer to your Additional Readers for a presentation titled 'Presentation of Data and Other Descriptive Measures:

1. Frequency tables
2. Histograms/Bar charts
3. Frequency polygons
4. Line graphs
5. Scatter plots
6. Stem and leaf plots
7. Box and whisker plots
8. Other plots/charts

Important! Refer to your **Additional Resources Folder** for the PowerPoint Presentation titled: "Presentation of Data and Other Descriptive Measures" by Roy Mayega. This presentation is a 'must-read' for you. In the presentation, look out for the following:

- What is the general lay-out of a frequency table? How does a frequency table that presents uni-variable data defer from one presenting multi-variable data?
- What are the different graphical methods and data presentation and when are they relevant? When for instance would you use a pie-chart, a histogram, a bar-graph and a line graph?

2.2.7 Extension Activities

Extension Activity 1: Discussion Forum Question

Statistical Methods make use of several measures of location and dispersion in describing data. Select one measure and describe its computation and application

Extension Activity 2: Self- Assessment Quiz

QUIZ 2.2.1

(Choose one correct option)

1. One of these statements is true about the arithmetic mean
    a. It is also called the geometric mean and is the commonest measure of location
    b. In it a valuable parameter in skewed arrays
    c. It is the middle value of a data array
    d. It is very sensitive to outliers

2. When the arithmetic mean is measured on a logarithmic scale, we obtain a transformed parameter called:
    a. The Geometric mean
    b. The Logarithmic mean
    c. The Exponential mean
    d. The Bayesian mean

3. In a distribution that is skewed to the right, the mean, mode and median are often unequal. One of these statements is true about such a distribution:
    a. The mode is often on the right of the median
    b. The Mean is often on the left of the median
    c. The Median is often on the left of the mean
    d. The median and mode are at the centre of the distribution, but the mean is not

4. Variance
    a. It is the sum of deviations of all individual observations from their arithmetic mean divided by the number of observations
    b. It is the sum of squared deviations of all individual observations from their mean, divided by the number of observations
    c. It is the sum of squares of all observations minus the squared sum of all observations
    d. It is the square root of the sum of squared deviations of all individual observations from their mean, divided by the number of observations

5. One of the following statements is true about graphical representations of data
    a. A box drawn against the Inter-quartile range, together with a line over the parts out of the range is called a “Stem and Leaf” plot
    b. The variation of one discrete continuous variable with another can be well represented by a line graph
    c. When comparing two distributions on the same graph, the best presentation is a histogram
    d. Line graphs and scatter plots are handy when presenting a distribution of random continuous variables

Dr. Mayega, in a field trial of a reproductive health intervention in secondary schools, obtained the following aggregate scores for 8 students in a post-instruction knowledge assessment test:

Student No.	1	2	3	4	5	6	7	8
Score	36	45	58	64	64	72	81	83

Use this information to answer questions 6 to 10

6. The Mean score is
    a. 62.9
    b. 79.2
    c. 64.0
    d. 69.2

7. The Median score is
    a. 58
    b. 61
    c. 88
    d. 64

8. The Modal Score is
    a. 72
    b. 81
    c. 64
    d. 45

9. The Inter-quartile Range is
    a. 36
    b. 47
    c. 27
    d. 14

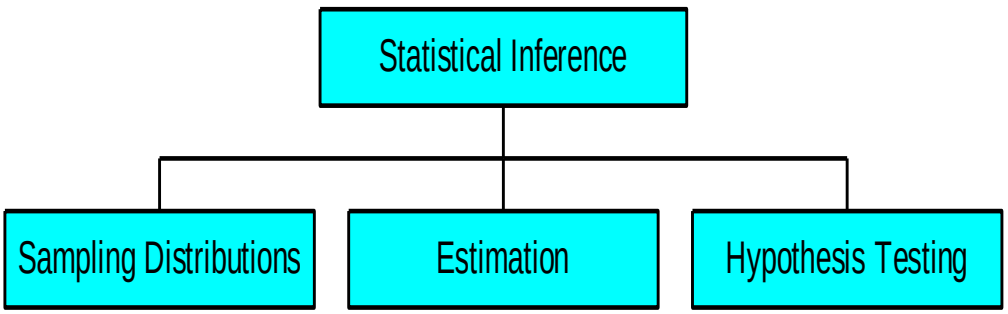
10. The variance is
    a. 269
    b. 407
    c. 16.4
    d. 32.8

2.3 Unit 3: INFERENTIAL STATISTICS I – PROBABILITY AND PROBABILITY DISTRIBUTIONS

2.3.1 Introduction to the Unit

Introduction to statistical Inference: The purpose of statistical inference is to enable you make decisions about population characteristics, using a sample. You start by describing the population, using 'descriptive statistics'. You then develop a hypothesis about differences in characteristics (between populations or sub-populations within the same population), so as to draw inference on 'association' if the sub-populations are found to differ significantly. Population characteristics are those things that distinguish or identify specific populations, for example the distribution of age, height, sex, etc. In statistics, we obtain information on a SAMPLE, by describing its characteristics; we then test hypotheses so as to generalize the sample parameters to the POPULATION. As you may have noted already, statistical inference has three branches: ***Probability distributions***, ***Estimation*** and ***Hypothesis testing***. These terms make up what we refer to as **inferential statistics.** While descriptive statistics is concerned with 'Univariable' attributes, inferential statistics is mainly concerned with comparisons between sub-populations and between two or more variables, hence bi-variate and multivariate analysis.

Figure 1: Branches of Statistical inference



In the first part of inferential statistics, we are going to look at the concept of probability as a basis for statistical inference. We shall look at the different probability distributions and the situations in which they apply. This will provide a basis for the second part of inferential statistics to be discussed later. Before we delve into estimation and hypothesis testing, it would also be important that we remind ourselves of the key issues in the probability and set theory.

2.3.2 Unit Outline

The following topics will be covered:

1. Background to Probability distributions and the concept of probability
2. The population distribution and the parametric approach
3. Sampling Distributions
4. Probability and the set theory

2.3.3 Instructional goal

The MPH STUDENT should be able to describe and contrast the different probability distributions and explain the situations in which they are applied in statistical inference.

2.3.4 Unit Objectives

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

1. Correctly explain the rationale for probability distributions
2. Contrast the population distribution with sampling distributions
3. Differentiate the various types of sampling distributions in terms of their characteristics and when they apply
4. Explain the key concepts in the set theory using a probability approach

REQUIRED READINGS

Wayne W Daniel (1998) Biostatistics: A Foundation for Analysis in the Health Sciences, 7th Edition (Published by John Wiley & Sons, Inc.)

Chapter 4: pages 83-99, 104-123.

Chapter 6: pages 161-166.

2.3.5 Time Frame

2 WEEKS

2.3.6 Content

Lesson 1: Background to the Concept of Probability and Probability Distributions and

Introduction: In statistics, we want to observe samples and infer our observations to populations. Most attributes in populations are distributed in a way that if we observe samples of given sizes, there is a likelihood or **probability** that particular proportions of persons will exhibit a particular characteristic. We can predict a range (or proportions or numbers) in which the majority of people lie, such that there is a very low probability that a person will lie beyond these ranges. People or populations that lie beyond this range are therefore 'beyond normal' and this is not likely to occur by chance. The basis for understanding these relationships is the Standard Normal Distribution. In this lesson, we shall discuss the basis for the 'normal distribution' and the rationale for the use of 'probabilities' or 'p-values' in bio-statistics.

Lesson Outline:

a. Background to probability distributions
b. The normal curve
c. The basis of probability distributions
d. Rationale for probability
e. Rationale for the p-value
f. Other parameters that follow probability distributions

Lesson Objectives:

By the end of this lesson, the MPH Officer should be able to:

1. Use the normal curve to illustrate the concept of probability in describing the random occurrence of characteristics in populations
2. Argue out the rationale and basis for probability distributions
3. Briefly explain the rationale and basis of the p-value and its use in bio-statistics
4. List at least 4 other parameters apart from the sample mean, that follow a probability distribution

a. Background to probability distributions

The Characteristics of a population are summarized using descriptive measures as discussed in the Unit on Descriptive Statistics. They include measures for numerical data (Means, Standard deviation) and measures for categorical data (Proportions). However, the attributes of a 'normal' population tend to take a predetermined trend or 'distribution'. Take height for instance; a population is likely to have a certain predetermined stratification of heights, with the majority of its people lying within a certain height range near the mean height. Very few people would be exceptionally tall or extremely short. Most people would be of medium height – neither too short nor too tall. The number of people in a particular height category would reduce as the heights increase or decrease away from the population mean.

The same characteristics are likely to be observed for weight, blood pressure, intelligence and any other measurable attributes. It is a characterizing feature of the 'normal population' and is therefore called the 'normal distribution'. When we know this predetermined tendency, we can calculate the probability that a person will have a certain height or a certain weight. If a person went out to outer space and came back with an altered systolic blood pressure, we can test whether the change in blood pressure is statistically significant or is still within the 'expected range' for people that did not go to outer space. We do this by collating the observed Blood pressure with the corresponding probability of its occurrence in the 'normal population'.

b. The normal curve

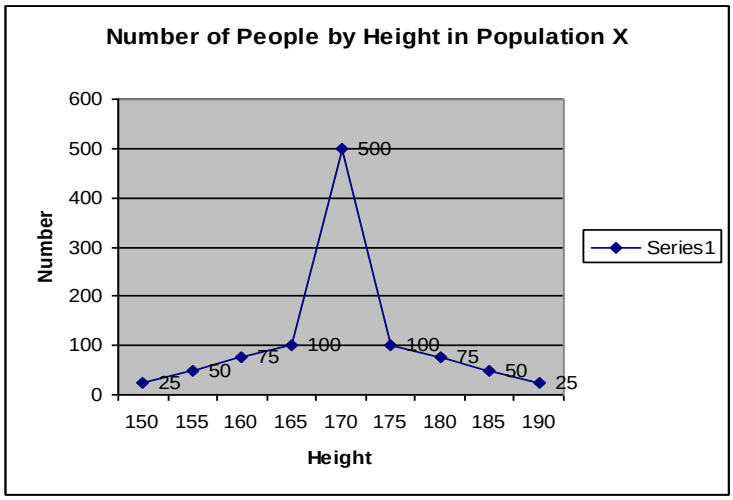
Assume we measure heights of the entire population of students in school X (1000 students in total) as indicated in the table below:

Example:

Height (cm)	Number	Percentage
150	25	2.5
155	50	5
160	75	7.5
165	100	10
170	500	40
175	100	10
180	75	7.5
185	50	5
190	25	2.5
Total	1000	100

If we plot these figures on a curve of height (x-axis) vs. number of people with a given height (y-axis), we are likely to obtain a bell shaped curve. The roughly bell shaped curve below is a very rudimentary example of a "normal distribution" curve. It represents the distribution of characteristics in a 'normal population'. What it simply portrays is that in a normal population, population characteristics are likely to be distributed in a pre-determined way, with the majority of the people lying near the mean. There are fewer and fewer people as we move away from the mean either towards extreme tallness or extreme shortness.

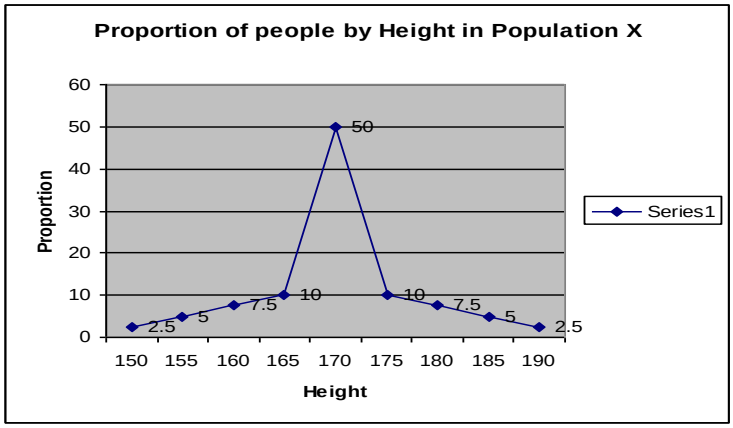
Illustration:



c. The basis of probability distributions

We can plot the above curve in terms of proportions, instead of absolute numbers i.e. proportions (or percentages) of people with a given height. It then becomes a probability distribution i.e. a graph of the probability of having given heights. Again, we see that there is a much higher probability of having a height near the population mean than having a height away from the mean. The figure below is an example.

Illustration:



Characteristics of the normal distribution:

- When data is normally distributed, approximately 70% of all people (68% to be exact) will lie within one SD (500+100+100)
- In this case, 1Standard Deviation is 5cm (i.e. at 165 or 175cm)
- Approximately 95% of all people lie between two Standard Deviations (500+100+100+75+75+50+50)
- In this case, 2 Standard Deviations is 10cm (i.e. 160 or 180cm)
- Therefore, only 5%, (or a probability, p, of 0.05) will lie outside this range
- This is the cut-off for p-values used in most statistical tests
- 98% of people lie within three standard deviations; we rarely use this cut off

d. Rationale for probability

We see that population characteristics are distributed using a pre-fixed model. Approximately 50% of people are at or near the mean value, 70% are within one Standard Deviation from the mean and 95% are within two Standard Deviations from the mean. We can know the probability that a person will have a height of say 180 cm. We can construct an entire table of probabilities that show the likelihood of occurrence of each height. We can do likewise for weight and many other population characteristics that follow a normal distribution – The resulting table is the same. We only need to use a standardized scale that recognizes height, weight, blood sugar etc. When we standardize the measure, we can then check from the probability tables to tell us the likelihood of occurrence of a given characteristic.

e. Rationale for the p-value

In most cases, when we conduct a study, we use a sample instead of a population. We measure a parameter in a sample. Suppose in the very same example above, we took only the students in Senior Six in the school and measured their mean height:

- There is a 50% probability (p=0.5) that their mean height will lie near 170, the mean of the entire population of students in the school
- And a 70% probability that their mean height will lie between 165-175cm (or within 1 standard deviation from the mean) (p=0.70)
- As well as a 95% probability that their mean height will lie between 160-180cm (or within 2 standard deviations from the mean) (p=0.95)

Looked at conversely:

- There is a 50% probability (p=0.5) that their mean height will not be near the school mean
- A 30% probability that their mean height will be less than 165 or more than 175cm (i.e. beyond 1 standard deviation from the mean) (p=0.30)

- A 5% probability that their mean height will be less than 160 or more than 180cm (i.e. beyond 2 standard deviations from the mean (p=0.05)

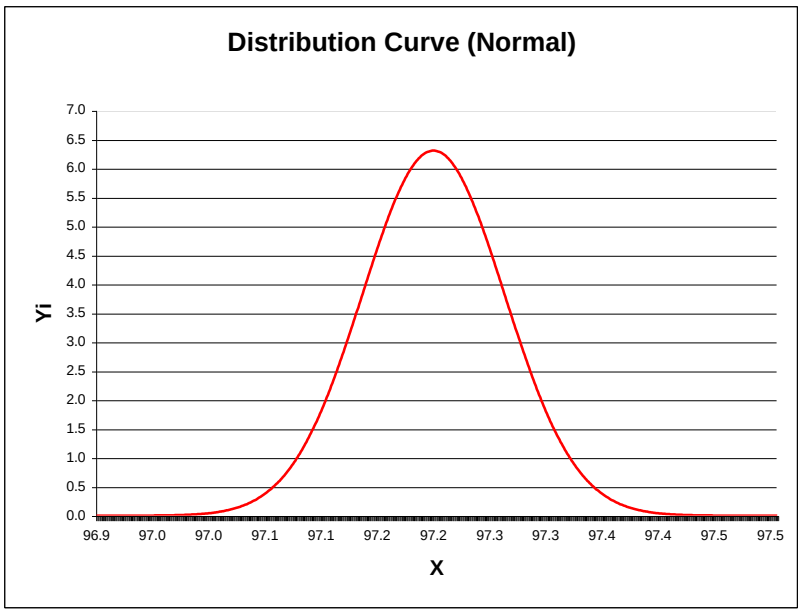
All these are based on probability or chance. The probability corresponding to non-occurrence is what is called the p-value

Interpretation: If our chance level is up to a probability of 5% or 0.05, then:

- Any p-value at 0.05 or more is expected, and occurs increasingly by chance
- Any p-value less than 0.05 is less and less likely to occur by chance

Therefore, if we find that the mean height of S6 students is more than 185cm, then the corresponding p-value is less than 0.05. Our minds are then alerted that the chances of this occurring in a normal population are less than 5% because our alert level has been set at p=0.05. Any p=value less than that is reported as not having occurred by chance, or is "statistically significant". We therefore conclude that this class is on average taller than expected, and we can then proceed to investigate why this particular cohort is so tall or so short.

Illustration: Changes in body temperature can be used to estimate when ovulation takes place. It is known that the average body temperature changes by up to a degree during ovulation. In order to measure the change, we need to know the mean basal body temperature. Suppose that for this purpose, we take the temperatures of an entire population of 100 girls in a hostel, all taken after ovulation, and just after waking up in the morning. Assuming that we obtain a mean basal temperature of 97.2, and a standard deviation of 0.2, we can construct a normal distribution curve as follows:



Using the curve: Supposing we then take the temperatures of a sub-sample of 50 girls during actual ovulation, and we find that it is over 0.4^0F higher (say 0.5^0F) (i.e. over 2 SDs); the probability of this occurring in "normal" (non-ovulating girls) is less than 0.05%. We conclude that ovulation is indeed associated with a rise in average basal temperature and that the difference is statistically significant (can't have occurred by chance). We re-enforce this by quoting a p-value corresponding to a temperature 0.5^0F higher than normal.

f. Other parameters that follow probability distributions

The mean of the population (with regard to height, weight, temperature etc) therefore follows a distribution called the “normal distribution”. In statistics, it is not only the population mean that follows a probability distribution. There are other parameters that do follow a normal distribution

- The mean of a sample follows what is called a t – distribution, which is very similar to the normal distribution, but, as we shall see later, is narrower and has a smaller standard deviation.
- The means of several samples follow a distribution called ‘the sampling distribution of means’
- Proportions of a given attribute follow a “normal distribution”
- The Standard Deviations of samples from a given population follow an “F- Distribution”
- Rates of occurrence of events follow a “Poisson” distribution
- Odds Ratios and Relative Risks follow a chi-square distribution

If we have a sample parameter of any of these we can determine the p-value, hence accept or reject set hypotheses on the likelihood of the occurrence of this particular parameter in a population. In the next two lessons, we shall discuss one by one the different types of probability distributions, starting with the population distribution in lesson 2 and then the sampling distributions in lesson 3. In lesson 3, we shall recap the standard normal distribution and discuss other sampling distributions commonly used in statistics.

Lesson 2: Probability Distributions: The Population Distribution and the Parametric Approach

Introduction: The population we discussed in the previous lesson is a 'normal population'. However, the *ideal population* has characteristics that are 'evenly' distributed. It is therefore called the 'population distribution. It is referred to as 'ideal' because it is rarely the case in real life. Before we go on to describe the characteristics of samples, let us see the characteristics of an ideal population. We also know that very rarely do we study an entire population. We instead use the characteristics of a sample to infer on the characteristics of a population. This is the basis for 'estimators'. The use of summary parameters (e.g. means, standard deviation and proportions) to describe the attributes of populations or samples is called the 'parametric approach'.

Lesson Outline:

a. Features of the population distribution
b. Estimators
c. The Parametric approach

Lesson Objectives:

By the end of this lesson, the MPH Officer should be able to:

1. Describe the key characteristics of an ideal population and the features of the population distribution
2. Select appropriate estimators for population parameters and explain their use in approximating sample parameters
3. Briefly explain the basis for the 'parametric approach'

a. Features of the population distribution

Unlike the sampling distributions, the population distribution is a uniform rectangular distribution. It assumes that characteristics are uniformly distributed in populations, and are not random. This idealistic population rarely occurs. Each height has the same number of people over the entire range; each weight has the same number of people in the entire range. While this may be true for a few characteristics, it is rarely the case for most random variables.

Population characteristics: As we have discussed over and over again, the characteristics of a population can be described using certain measures called **population parameters**.

The most important parameters for continuous variables are:

- the population mean (often denoted as μ)
- the population standard deviation σ and
- the difference between population means $(\mu_1 - \mu_2)$

The most important parameters for categorical variables are:

- the population proportion (proportion of people with a particular attribute) denoted as P
- the difference between population proportions $P_1 - P_2$

b. Estimators

Statistics is about measuring **sample parameters** and using them as **estimators** for the population parameters.

For continuous variables:

- the sample mean $(\bar{x})$ is an estimator of the population mean (μ)
- the sample standard deviation (s) is an estimator of the population Standard Deviation (σ)
- the difference between sample means $(\bar{x}_1 - \bar{x}_2)$ is an estimator of the difference between population means $(\mu_1 - \mu_2)$

For categorical variables:

- the sample proportion (p) is an estimator of the population proportion (P)
- the difference between sample proportions $(p_1 - p_2)$ is an estimator of the difference between population proportions $(P_1 - P_2)$

Illustration: Relationship between population and sample parameters

	Population	Population parameter	Sample	Estimator
1.	Population mean	(μ)	Sample mean	$(\bar{x})$
2.	Population SD	(σ)	Sample SD	(s)
3.	Difference between population means	$(\mu_1 - \mu_2)$	Difference between sample means	$(\bar{x}_1 - \bar{x}_2)$
				$se = (\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$
4.	Population proportion	(P)	Sample proportion	(p)
5.	Difference between population proportions	$(P_1 - P_2)$	Difference between sample proportions	$(p_1 - p_2)$
				$se = \sqrt{\frac{p_1.(1-p_1)}{n_1}} + \sqrt{\frac{p_2.(1-p_2)}{n_2}}$

c. The Parametric approach

The use of all these measures that summarize characteristics of populations and samples is called the parametric approach; the summary measures are therefore called parameters. The most commonly used parameters therefore are: The mean and the standard deviation for continuous variables, and the proportions for categorical variables.

Lesson 3: Probability Distributions: Sampling Distributions

Introduction: As we have been emphasizing and re-emphasizing, it is not always possible to study an entire population. We therefore often study a sample and 'extrapolate' or 'generalize' the findings to the general population. We also saw in lesson 1 that the characteristics of normal populations tend to follow certain predetermined distributions, and that because they follow a fairly consistent pattern, we are able to construct a probability distribution of the proportion of people or items that are likely to have a certain measurement of say height, weight, basal temperature etc. It enables us to predict, with a fairly good level of accuracy, the likelihood that a certain person, taken at random from the population, will have a certain measurement. It also enables us to distinguish the 'abnormal' or 'statistically different' people.

Unlike the ideal population we saw in the previous lesson, samples of populations tend not to have uniformly distributed characteristics. They instead follow characteristic trends based on probability of occurrence of the different measures in the range. A vivid illustration was shown to you in lesson 1in which we used hypothetical heights. These relationships are what are described as the *'sampling distribution'*. The sampling distribution is therefore a mathematical relationship that assigns a probability of occurrence to the different values of a given variable.
The variable can be:

- Height
- Weight
- Basal body temperature etc.

ACTIVITY: Think of other attributes that cab follow a probability distribution.

We therefore have predetermined distributions that enable us to predict with certainty, the attributes of populations based on samples. These are called sampling distributions. In the next sub lessons, we shall one by one consider these distributions. A sampling distribution can be displayed in form of a table giving the random values and their associated probabilities, or it can be expressed as a mathematical formula. Additionally, it can be displayed graphically. Graphical representation helps us to visualize the shape of the distribution.

Some common sampling distributions: The common sampling distributions are:

- The standard normal distribution
- The sampling distribution of the mean
- The students' t-distribution
- The binomial distribution
- The F-distribution
- The Poisson distribution
- The Chi-squared (x^2) distribution

It is worth noting that sampling distributions do not apply to the mean alone: we also have sampling distributions of **proportions, medians, correlations, rates** (Poisson), **variances** and **ratios of variances** (F-distribution). These distributions and their applications will be the subject of the next few sections. The chi-squared distribution will be discussed in another unit.

Lesson Outline:

Sub-Lesson 3a: The Standard Normal Distribution
Sub-Lesson 3b: The Sampling Distribution of the mean
Sub-Lesson 3c: The t – Distribution
Sub-Lesson 3d: The Sampling Distribution of Proportions
Sub-Lesson 3e: The Binomial Distribution
Sub-Lesson 3f: Introducing other probability distributions

Lesson Objectives:

By the end of this lesson, the MPH Officer should be able to:

1. Illustrate the sampling distribution of the mean and its use in answering questions about populations
2. Evaluate the t - distribution and contrast its applicability from other sampling distributions
3. Select appropriate parameters to describe the sampling distribution of proportions
4. Describe the binomial distribution and its applicability in dealing with categorical data
5. Name and briefly describe the use of other distributions that are important in statistical analysis

Sub-Lesson 3a: The Standard Normal Distribution

a. What is it?

The normal, Gaussian or "bell-shaped" distribution is one of the ***sampling distributions*** to be discussed in this lesson. As you read this section, it is important for you to refer back to the information in Lesson One. In that lesson, we used the Standard Normal Distribution to illustrate the concept of probability. It is the cornerstone of most of the methods of estimation and hypothesis testing. It is a continuous sampling distribution. Many random characteristics of the general **population**, such as the distribution of birth-weights or BP in the general population, tend to approximately follow a normal distribution. In addition, many random variables that are not themselves normal approximately follow a normal distribution when summed many times. The normal distribution is generally more convenient to work with than any of the other distributions, particularly in hypothesis testing. Thus, if an accurate normal approximation to some other distribution can be found, then we will often use it.

Example 2.3.6: The number of neutrophils in a sample of 5 white blood cells is not normally distributed, but the number in a sample of 100 white blood cells is very close to being normal.

Characteristics of the normal distribution: The normal distribution has the following properties:

- its probability function follows a bell-shaped curve,
- the curve is symmetric about the mean μ, which is the most frequently occurring value
- the variance of the distribution is σ^2 and its standard deviation is σ
- about 68% of the observations lie within 1 standard deviation for the mean
- about 95% of observations lie within two standard deviations and 99.7% of observations lie within 3 standard deviations

It can be mathematically shown that the mean μ and σ^2 are respectively the expected value and variance of the normal distribution. These two parameters are indeed the **population parameters**. Once these two parameters are known, the normal distribution is defined. Thus, the entire shape of the normal distribution is determined by the two parameters.

Remember!
Most population characteristics that are continuous in nature follow a normal distribution, with a mean **μ** and a variance **σ²**. These as you recall are known as the **population parameters**

b. Standardizing the normal distribution

The different observations in an array of normally distributed data can be represented in a normal distribution curve. This curve is the same for height, weight, basal temperature and all other characteristics that follow a normal distribution. However, the horizontal scale of this curve is different for height, weight, temperature, systolic blood pressure, because of the different units used to measure them.

For example, the heights of adult people may range from 140cm to 190cm, while normal blood sugar levels may range from 80 to 120mg/l, and weights of normal adults may range from 40kg to 120kg. As you can see, all these are characteristics use different scales. Yet they all follow a similar curve.

To enable the use of a common table of probabilities, we need to standardize the different characteristics (height, weight etc) so that they can fit on the same horizontal scale. This is called standardisation.

The scale derived by standardisation of a normally distributed attribute is called the ***z-scale***. On this scale, when the population mean is standardized, it equals to zero $(\mu = 0)$. Since it is at the centre of the data, there are negative values to the left of the mean and positive values to the right. Using this scale, we can compare the distributions of data on different variables e.g. weight, height etc (that are otherwise measured on different scales), on a common scale.

The process of converting a particular variable to the ***z-scale*** is called ***standardization*** and the value we obtain after standardizing is called the ***z-statistic***. The resulting distribution is called the ***standard normal distribution*** or the ***z-distribution***. To standardize a particular variable x_i , we use the following formula:

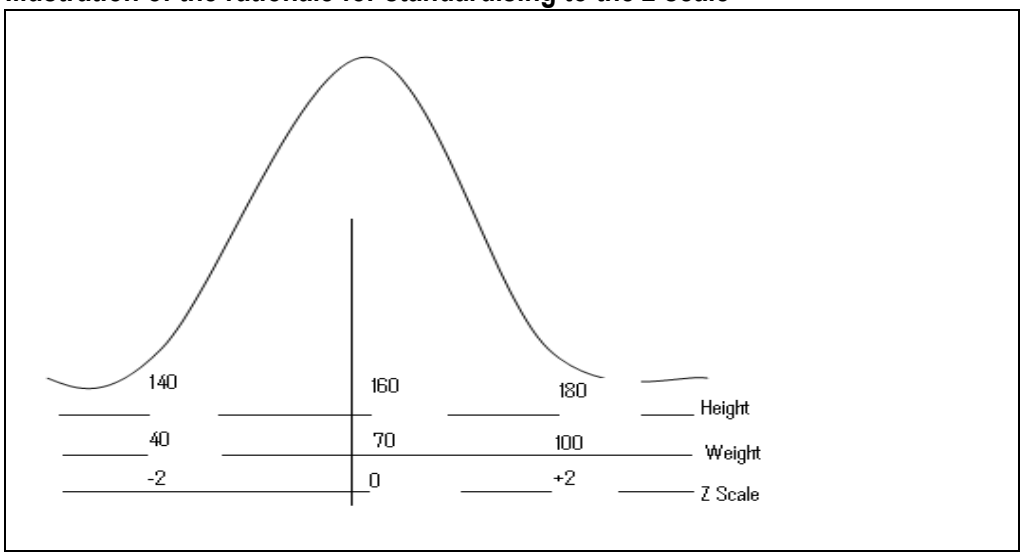
$$z = \frac{Variable - Mean}{S\tan dardDeviation}$$

Or: $$z = \frac{Variable - Mean}{\sqrt{Variance}}$$

i.e.: $$z = \frac{x_i - \mu}{\sigma}$$

As you can see, the z-value is a ratio of the difference between an observation and its mean $(x_i - x)$ to the Standard Deviation. The z-value is therefore called the ***critical ratio***.

Illustration of the rationale for standardising to the z-scale



Sub-Lesson 3b: The Sampling Distribution of the Mean

The normal distribution above applies to a population. However, in practice, we rarely study entire populations and we instead study samples of the population. When we take repeated samples from a given population and compute the sample parameters (e.g. the mean), the means of the different samples that we take tend to follow a 'near normal' distribution when plotted. This distribution formed by taking repeated measurements is known as 'the sampling distribution of means'. The sampling distribution of the mean is a theoretical distribution that borrows the features of a normal distribution.

a. Deriving the sampling distribution of the mean

- We start with a particular population.
- We draw several samples of size n till we have exhausted all the possible samples of than size that we can draw from that population
- We compute the individual means $(\overline{x})$ of all the samples
- All the possible means $(\overline{x}_i)$ that we obtain can be plotted into a probability distribution
- This distribution has a mean $\overline{\overline{X}}$ (the mean of many sample means) and a standard deviation $\sigma_{\overline{X}}$ (The standard deviation of the means of many samples from their overall mean) Thus the shape of this distribution is determined by these two parameters.

The sampling distribution as an estimator of the population parameters

When we compare it with the derivative population, we observe the following:

- The mean of the sample means $\overline{\overline{X}}$ is similar to the population mean (μ)
- However, interestingly, the sample standard deviation $\sigma_{\overline{X}}$ is in fact not equal to the population standard deviation (σ); it is instead smaller.

Note therefore:
While the mean of the sample means $(\bar{x})$ approximates to the population mean μ, the standard deviation of the sample means is slightly less than that of the population, the reason being that using sample means averages some of the more extreme values observed in a population. The resulting distribution is a bell-shaped distribution, very similar to the normal distribution, but in fact narrower.

The standard deviation of the sampling distribution, which has reduced as compared to the population parameter, as a result of the sampling is called the **standard error of the mean**. The standard error of the mean is in fact given by the population standard deviation divided by the square root of the sample size used to draw the samples **(n)**.

se: = $\frac{\text{Population Standard Deviation}}{\text{Square Root of Sample Size}}$

se: $\frac{\sigma^2}{\sqrt{n}}$

NB: These facts are true if the repeated samples are drawn ***with replacement***

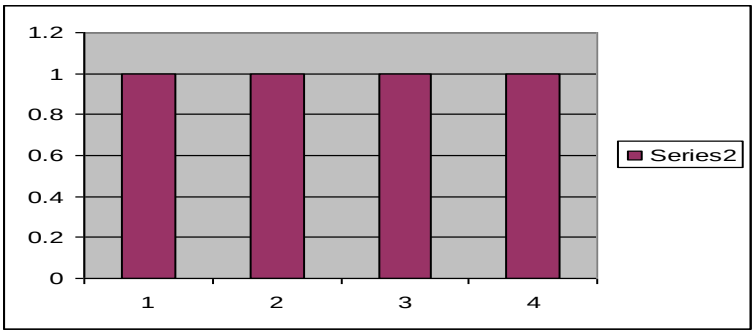
Example 2.3.7: Illustration

- Suppose there is a population of size N=4
- In it there is a variable X, which is a random variable
- The values of X (x_i) in this population are **1, 2, 3** and **4**
- We assume they are uniformly distributed

When we compute the population characteristics, the summary measures (or parameters) are:

Mean: $\mu = \frac{\sum_{i=1}^{n} x_i}{N}$ **= 2.5;** **SD:** $\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$ **= 1.12**

The resulting population distribution is as below:



Now, let's derive a sampling distribution:
We take all the possible samples of size $(n = 2)$: The maximum number of samples we can have in this case is 16 samples.

Therefore, we can obtain 16 sample means as shown below:

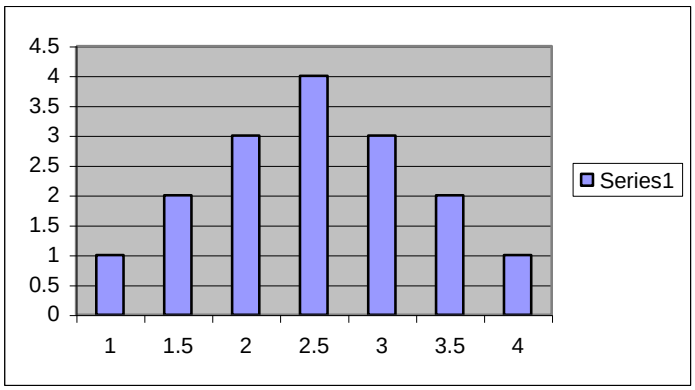
16 Samples

1st Observation	2nd Observation			
	1	**2**	**3**	**4**
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

16 Sample Means $(\bar{x}_i)$

1st Observation	2nd Observation			
	1	**2**	**3**	**4**
1	1.0	1.5	2.0	2.5
2	1.5	2.0	2.5	3.0
3	2.0	2.5	3.0	3.5
4	2.5	3.0	3.5	4.0

The resulting sampling distribution (of sample means) is as shown in the figure below:



We then compute the characteristics of the sample means; the summary measures are:

Mean: $\bar{X} = \dfrac{\sum_{i=1}^{n} \bar{x}_i}{N}$ = $\dfrac{1.0 + 1.5 + \ldots.. + 4.0}{16}$ = **2.5**

SD: $\sigma_{\bar{x}} = \sqrt{\dfrac{\sum(\bar{x}_i - \bar{X})}{N}}$ = $\dfrac{(1.0 - 2.5)^2 + (1.5 - 2.5)^2 + \ldots.. + (4.0 - 2.5)^2}{16}$ = **0.79**

Observation:
You will see that the mean of the sample means is equal to the population mean. However, we observe clearly here that the ***population standard deviation is not the same as the SD of the sample means***. The SD of the sample means is smaller and is given by the population standard deviation divided by the square root of the sample size used in drawing the repeated samples.

$$\frac{(\sigma)}{\sqrt{n}} = \frac{1.12}{\sqrt{2}} = \frac{1.12}{1.414} = 0.79$$

It is in fact less than the population SD. It is called the Standard Error of the Mean, designated (se)

b. Standardizing the Sampling distribution of the mean

As we did for the normal distribution, we can apply the same formula to standardize the distribution of the sample means (However, note that the SD of the sampling distribution is now equal to the population SD, but rather, the population SD divided by the root of the sample size of the samples drawn:

$$z = \frac{Variable - Mean}{S\tan dardDeviation}$$

Or: $$z = \frac{Variable - Mean}{S\tan dardError}$$

i.e.: $$z = \frac{(x_i - \bar{x})}{\sigma/\sqrt{n}}$$

c. The Central Limit Theorem

This theory is very important because many of the distributions encountered in daily life are not normal. In such cases the central limit theorem can often be applied. This will allow us to perform statistical inferences based on the approximate normality of the sample mean, despite the non-normality of the distribution of individual observations.

Let $y_1, y_2, ..., y_n$ be a random sample from some population with mean μ and variance σ^2. Then for large n, the sample mean $\bar{x}$ ~ N (μ, $\sigma/\sqrt{n}$) even if the underlying distribution of individual observations in the population is not normal.

The Central limit Theorem

If random samples of size n are repeatedly drawn from any population, with mean μ and variance σ^2, then for a large n (i.e. n> 30), the distribution of the sample means $\bar{x}_i$ will be approximately normal, with mean μ and variance σ^2/n [Standard Deviation of $\sigma/\sqrt{n}$]. This is called the **Central Limit Theorem**.

d. The sampling distribution of the difference between two means

If both samples are random and independent, and both populations are normal or sample sizes are 30 and above, the distribution of the difference between two sample means is approximately normal with the following parameters:

The difference between sample means: $(\bar{x}_1 - \bar{x}_2)$ approximates to the difference between population means $(\mu_1 - \mu_2)$

The variance of the difference $(\sigma^2_{x_1 - x_2})$ is given by: $\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$

The standard error of the difference $(\sigma_{x_1 - x_2})$ is given by: $\frac{\sigma}{\sqrt{n_1}} + \frac{\sigma}{\sqrt{n_2}}$

Sub-Lesson 3c: The Student's t – Distribution

a. The t-distribution, when the population parameters are known:

The sampling distribution of the mean assumes that we take several samples in any given study, and compute 'the mean of means'. However, because of logistical constraints, we cannot do this in routine research – we do not have the resources to take several repeated samples and make measurements. In practice, we usually take only one sample. When we take one sample, the sample parameter (or mean in this case) is likely to be different from the population mean (i.e. there is an error). The distribution of a single sample is therefore less likely to be the same as the normal distribution. The distribution of a single sample is also dependent on the size of that sample.

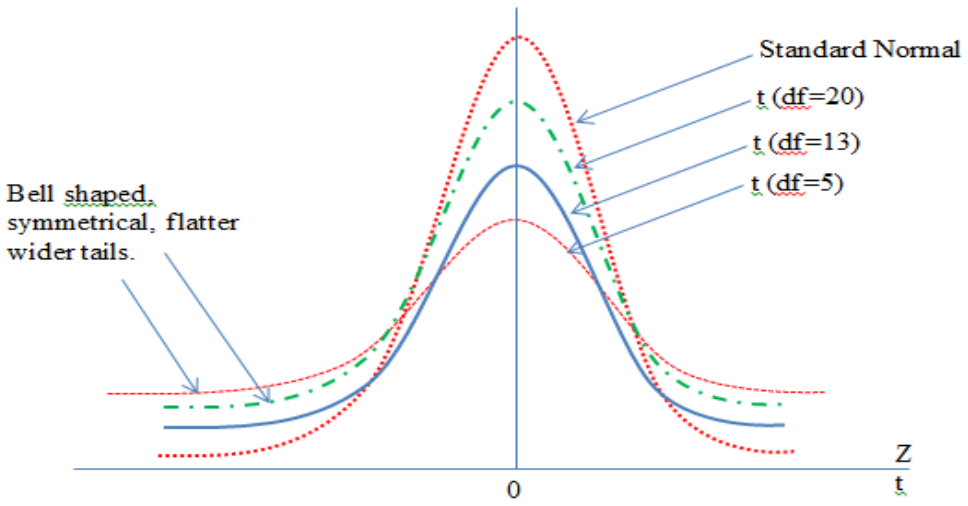
Following the description of the normal distribution by Gauss, a mathematician who at the time of writing it was in prison, had studied the normal distribution and discovered that in small populations (with few numbers), the shape of the curve changes according to the number of items in the population. Because he wanted to disguise himself, he published the discovery and called the distribution a "Student's t-distribution". It is now commonly known as the t-distribution. Therefore, there is no unique t-distribution but instead a family of t-distributions. This family of distributions is indexed by three parameters, namely

- the mean value of the population, μ
- the variance of the population, σ^2; and
- the degrees of freedom (df) in the population given by $(n-1)$

It is therefore a continuous probability distribution. The application of the t-distribution is very similar to that of the normal distribution except that when the sample size of study is small (usually anything below 40), it is advisable to use the t-distribution. The t-distribution has the following properties:

- the shape of the probability function changes with the sample size,
- the shape stabilizes to a bell-shaped curve as the sample size increases
- the curve is symmetric about the mean (μ), the most frequently occurring value
- the variance of the distribution is σ^2

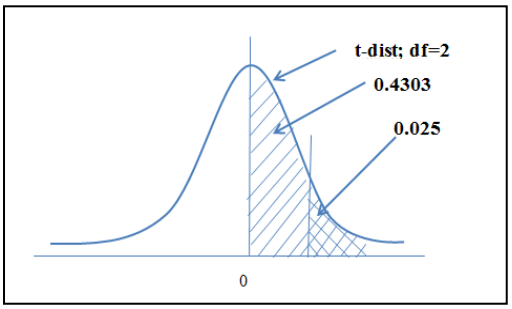
Illustration:



The t-table: The t-table differs slightly from the Z- table in that it also has degrees of freedom.

Illustration: (Upper tail area)

Assume: n = 3
df = $(n-1)=2$
α = 0.05
$\alpha/2$ = 0.025



	Upper tail area		
p-value	0.10	0.05	0.025
df			
1	3.078	6.318	12.706
2	1.886	2.920	**4.303**
3	1.638	2.353	3.182

The reverse situation is true for the lower tail area.

b. The t-distribution, when the population parameters are not known

When planning to use the sampling distribution of the mean, we assume that the population standard deviation is known. The mean of the sample means $(\bar{X})$ approximates to the population mean μ and the standard deviation of the sample means $(\bar{x}_i)$ ***(standard error of the mean)*** is computed from the population standard deviation as $\sigma/\sqrt{n}$ **.However**, in real life,

- we do not readily know the population SD (σ)
- we do not take may samples; we often take one representative sample

In these cases, we can instead use the sample standard deviation of the actual sample we are using, to calculate the standard error of the mean [instead of $\sigma/\sqrt{n}$ **,** we use $s/\sqrt{n}$. When we use the sample standard deviation (s) instead of the population parameter (σ), the resulting distribution of the mean actually does not follow the sampling distribution of the mean. It follows a special distribution known as the students' t-distribution. It is bell shaped, like a normal distribution, but is narrower in dispersion, than both the sampling distribution of means and the normal distribution. It is a more cumbersome distribution to use, because it is affected by ***degrees of freedom.*** As already mentioned earlier, it is affected by the sample size and each sample size has its own distribution – hence a family of distributions. However, the fact that we do not need prior knowledge of population parameters makes this distribution one of the most frequently used distributions to test hypotheses in the practice of statistical inference.

Sub-Lesson 3d: Dealing with Categorical data – The Sampling Distribution of Proportions

So far, we have been dealing with the sampling distribution of a mean. Categorical variables like sex can be counted. In such cases, we count the proportion of the population having a particular characteristic. There are two possible outcomes in the sample and in the general population: those who possess and those who do not possess a particular characteristic. This can be expressed as a proportion.

$p_S = X/n$ = Number of successes / Sample size

a. The sampling distribution of a proportion

It approximates to a **normal distribution**, determined by two parameters:

P = the population proportion

σ_P = the standard deviation of the population proportion

However, for most purposes, the population parameters are not known, and we use the sample parameters to estimate the population parameters:

The sample proportion (p) approximates to the population proportion P

The sample standard deviation of proportions (s_p) approximates to the standard deviation of the population proportions (σ_P)

Like the standard deviation of the sample mean which is slightly less that the true population SD the standard deviation of the sample proportion is also slightly lower than the SD of the population proportions and is given by the formula:

$$s_p = \sqrt{\frac{p.(1-p)}{n}}$$

b. Standardizing the sampling distribution of the proportion

We can standardize the sampling distribution of a proportion so that we can use it to make estimates on the z-scale; the critical ratio (z-statistic) is in this case given by the formula:

$$z \equiv \frac{(p_s - P)}{\sqrt{\frac{p.(1-p)}{n}}}$$

c. The sampling distribution of the difference between two proportions

If both samples are random and independent, and both populations are normal or sample sizes are 30 and above, the distribution of the difference between two sample proportions is approximately normal with the following parameters:

The difference between sample proportions $(p_1 - p_2)$ approximates to the difference between population proportions $(P_1 - P_2)$

The variance of the difference $s_{(p_1-p_2)}$ is given by: $\sqrt{\frac{p_1.(1-p_1)}{n_1}} + \sqrt{\frac{p_2.(1-p_2)}{n_2}}$

Lesson 3e: The Binomial Distribution

Introduction: It is the sampling distribution of proportions in which the outcome has two categorical outcomes: say success or failure. When we take repeated samples of size n from a population, the proportion of items with a particular attribute (X/n) follows a normal distribution. Since it is a distribution of proportions of one of two categorical options, it called the binomial distribution.

Characteristics: This is a sampling distribution with the following properties:

- There is a sample of N independent trials,
- Each trial can have only two possible outcomes, which is often denoted by either "success" or "failure",
- The probability of success or failure is assumed to be constant at each trial

Example 2.3.8: One of the most common laboratory tests performed on any routine medical examination is a blood test. The two main aspects to a blood test are: counting the number of white blood cells, differentiating the white blood cells that do exist into five categories, namely, neutrophils, lymphocytes, monocytes, eosinophils, and basophils (referred to as the "differential"). Both the white cell count and the differential are extensively used for clinical diagnosis. It is usually of interest to know the number of neutrophils, (k) out of every 100 white cells. For our purpose however, rather than out of 100 cells, let us consider out of 5 cells. What is the probability that the second and the fifth cells considered will be neutrophils and the remaining cells non-neutrophils, given the probability that any one cell being neutrophils is 0.6.

If the event of finding a neutrophils is denoted by S (success) and that of finding a non-neutrophils is denoted by F (failure), then the question being asked is: What is the probability of the outcome FSFFS = Pr (FSFFS)? Note that the order here matters; therefore you can only have one such ordering! Since the probabilities of success and failure are respectively 0.6 and 0.4, and the outcomes for different cells are presumed to be independent, then:

$$\begin{aligned} \Pr\{FSFFS\} &= \Pr\{F\} \times \Pr\{S\} \times \Pr\{F\} \times \Pr\{F\} \times \Pr\{S\} \\ &= (0.6)^2\,(0.4)^3 \\ &= 0.023 \end{aligned}$$

Exercise 2.3.1: What is the probability that two of the 5 white blood cells, will be neutrophils? (Here note that the order does not matter). We need to consider all the possible combinations where there are only 2 neutrophil cells. These are given by:

$${}_5C_2\,(p)^2(q)^3 = \binom{n}{k}(p)^k(q)^{n-k}$$

$$= \binom{5}{2}(0.6)^2 0.4^3$$

$$= \left(\frac{5!}{3!2!}\right)(0.6)^2 0.4^3$$

= 0.2304

It can be shown that the expected value of the binomial distribution is given by

E(X) = μ
= x.Pr(X=x), where the summation is over all possible values of X.
= n.p

It can also be shown that the variance the of binomial distribution is given by

Var(X) = σ^2
= $E(X - \bar{\mu})^2$
= $x_i Pr(x_i) - \mu^2$
= n.p.q

Lesson 3e: Introduction to other Probability Distributions

The F-distribution: Some times in statistical analysis, we may need to compare the difference between the means of more than two datasets, to assess whether they are homogenous. In such cases, we can do so by comparing variability among the data sets. We compute variances within groups (S_1^2) and the variances across groups (S_2^2) and their ratios. It can be shown that the ratio of variances follows an F-distribution under the null hypothesis that $(S_1^2 = S_2^2)$. This distribution of the variance ratio (S_1^2 / S_2^2) was extensively studied by the statisticians R.S. Fisher and G. Snedecor. Like the t-distribution, there is no unique F distribution but instead a family of F-distributions. This family of distributions is indexed by two parameters termed as the numerator and denominator degrees of freedom (df), respectively. Specifically, if the sample sizes of the first and second samples are n_1 and n_2 respectively, then the variance ratio follows an F distribution with $(n_1 - 1)$ (numerator df) and $(n_2 - 1)$ (denominator df)), which is denoted by $F_{(n_1-1),(n_2-1)}$. The F-distribution is generally positively skewed, with the skewed ness dependent on the relative sizes of the two degrees of freedom.

The Poisson distribution: The Poisson distribution is used to determine the probability of occurrence of rare events. It is a distribution of rates. It is a discrete distribution that is applicable when the outcome is the number of times an event occurs. It gives the probability that an outcome occurs a specified number of times when the number of trials is large, and the probability of one occurrence is small.

Practical Uses: Planning of the number of beds in an intensive care unit or district hospital

1. Planning number of ambulances needed
2. Projecting the number of cells in a given volume of fluid
3. Number of bacterial colonies in a certain amount of medium
4. Emission of radio-active particles from a specified amount of radioactive material

Consider a random variable, representing the number of times an even occurs in a given time or space interval. The probability of x occurrences is given by:

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

λ ≈ Value of both mean and variance of the Poisson distribution
e ≈ The base of the natural logarithm (2.718)

[We shall learn more about this distribution in the module Applied Epidemiology II – Analysis of cohort Studies].

Lesson 4: Probability and the Set Theory

Definition: Probability is the degree of certainty or uncertainty of occurrence of an event.
It: Provides a foundation for statistical inference
Is measured by either a number from 0 to 1 or as a proportion/fraction
Probability is closely related to the set theory.

Lesson Outline:

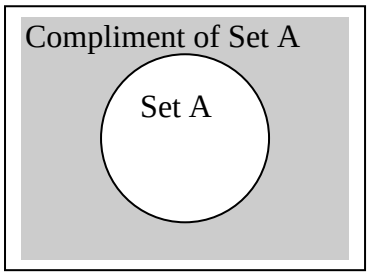
a. Basic concepts in set and probability theory
b. Combined probabilities
c. Baye's Theorem

Lesson Objectives:

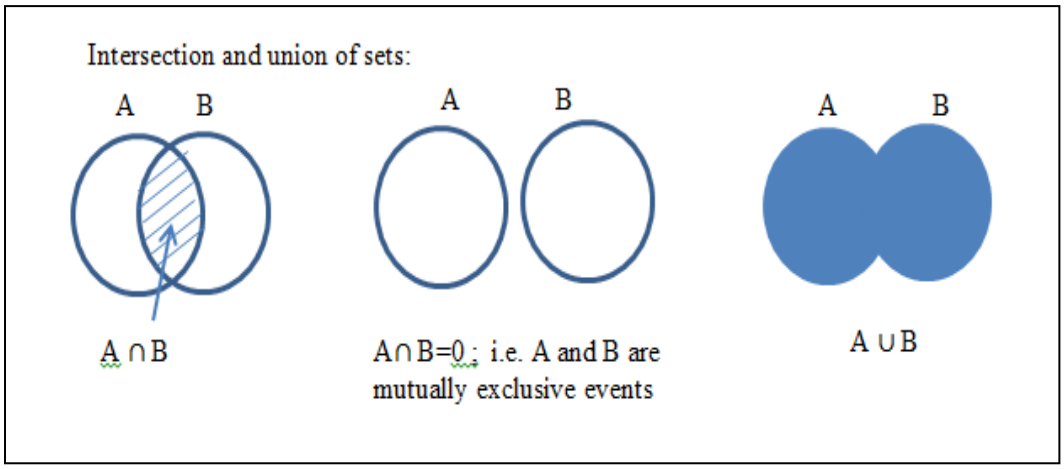
The MPH Officer should be able to:
1. Explain the key concepts in the probability theory.
2. Distinguish between mutually exclusive and independent events.
3. Distinguish between joint, marginal and conditional probability.
4. Derive and equation that summarises the multiplication and addition rules of probability.

a. Basic concepts in set and probability theory

Members of a set: The members of a set are the elements, items or events. If we pick a few of the members, them we have a sub-set. If we get a particular set of events (set A), that is complete and is not a subset, other elements outside the set are called the compliment of set A, denoted A^C or $\bar{A}$.



Intersection and union of sets: Sets can intersect and can be added (union) as illustrated below:



Probability views: There are two paradigms in probability:

The Objective paradigm: Assumes equally likely events that are of a long run nature, and not personal beliefs. What is observed is the same for all observers.

The Subjective paradigm: It involves personal belief and judgment, which differs between observers. It does not rely on repeatability of events.

Common Terms:

Universal set: A set containing all the possible elements. It is denoted { } or □

Partitioning a set: We can partition a universal set into subsets of mutually exclusive events

Sample space: Set of all possible elements, outcomes or events in a given intervention

Mutually exclusive events: The occurrence of one event excludes the occurrence of another i.e. the two cannot occur in the same subject at the same time: one cannot be male if they are female

Independent events: Occurrence of an event is not dependent on another, though the two can occur together e.g. Being female and being an alcoholic

Dependent events: Occurrence of an event depends on another e.g. lung cancer as a result of smoking

Complementary events: These are automatically mutually exclusive

Intersection of events: Simultaneous occurrence of two events. These may be independent or dependent.

An exhaustive set: It includes all possible options e.g. an exhaustive set of all religions

Deriving probability:

If an event can occur in **N** mutually exclusive, exhaustive and equally likely ways, and ***m*** of these have a trait **E**, then the probability of occurrence of E is ***m*/N**.

Example: Tossing a dice can result in six equally likely outcomes, and one of these outcomes is associated with a face with 2-dots. This face appears on only one side of the dice. The probability of getting a face with 2-dots therefore is 1/6. This probability may vary slightly if a few more trials are made. However, as more trials are made, this probability stabilizes at 1/6.

The probability of an event is defined as the relative frequency of a set of outcomes of the event, over an indefinitely large (or infinite) number of trials. If the process above is repeated many times (different samples of size n), and some resulting event E occurs m times, then the probability of event E will approximate to m/n i.e**. P (E) ≈ *m*/n**

Understanding probability is essential in the calculation and interpretation of p-values in statistical hypothesis testing (to be seen later).

Notation:	1. The symbol { } is used to denote the phrase "the event" 2. {A ∪ B} is the event that either A or B or both A and B occur 3. {A ∩ B} is the event that both A and B occur at the same time 4. A' (or AC or Ā) is the likelihood event that A does not occur. It is sometimes referred to as the compliment of A. Notice the Pr {A'} = 1 – Pr {A}, since A' occurs when A does not occur.

Properties of P (E) = m/N: It has the following properties:

1. It is always a non negative number $(PE_i) \geq 0$; Where E_i is an event that is mutually exclusive.
2. It ranges between 0 and 1 i.e. $0 \leq (PE_i) \geq 1$
3. It is exhaustive: $[(PE_1) + (PE_2) + (PE_3) + \ldots.. + (PE_{xn})] = 1$

b. Combined probabilities

1. Exhaustiveness: For one set of events E_X: $(PE_1) + (PE_2) + (PE_3) + \ldots.. + (PE_{xn}) = 1$

2. Two mutually exclusive events: The probability of occurrence of either is the sum of the sum of the individual probabilities

P (Occurrence of either) or P (E_i or E_j) = P (E_i) + P (E_j)
For example: If P (being female) is 0.5
P (being male) is 0.5
Since they are mutually exclusive, P (A ∩ B) = 0
And: P (A ∪ B) = P (A) + P (B)

Now, the probability of being either one of them P (being male or female) = 0.5 + 0.5 = 1. If this adds up to one, then it is complementary. However it may not add up to one e.g.

Example 2.1.X: If all professions are considered as skilled, unskilled, semi skilled and unclassifiable skill, and

P (skilled) = 0.25
P (unskilled) = 0.25
P (semi skilled) = 0.25
P (unclassifiable skill) = 0.25

Then the probability of being skilled or semiskilled would be 0.25 + 0.25 = 0.5
The probability of being skilled or unskilled or semi skilled or unclassifiable is 0.25+0.25+0.25+0.25 = 1
This further confirms that:

1. For one set of events E_X: $(PE_1) + (PE_2) + (PE_3) + \ldots.. + (PE_{xn}) = 1$
2. $0 \leq (PE_i) \geq 1$

4. **Two independent events:** For two independent events (A and B), the probability of being BOTH A and B is the product of the two probabilities and is denoted as the intersection of the two sets.
   = P (Being A) X P (Being B)
   i.e. P (A ∩ B) = P(A) X P(B)
   e.g. the probability of being male AND having malaria is an intersection of the two sets
   = P (Male) X P (Malaria)

5. Complementary events: If two events are complementary, then they are mutually exclusive. Therefore, if we have set A, then the complementary set is A' or A^C or $\bar{A}$
Therefore: $P(A) = 1 - P(A^C)$
And: $P(A) + P(B) = 1$

Contingency tables: The multiplication rule
Example: Times taking cocaine by gender

Times taking cocaine	Male (M)	Female (F)	
X	A	B	**(a+b)**
Y	C	D	**(c+d)**
Z	E	F	**(e+f)**
	(a+c+e)	**(b+d+f)**	**N**

1. Marginal Probability: If any of the sub-totals is used as a numerator, and the grand total N as the denominator.
e.g. Probability of being male (regardless of characteristic) = **(a+c+e)/N**

2. Joint Probability: If any of the individual cells is used against the grand total
e.g. Probability of being a *male* **AND** *taking cocaine X times* = a/N.
As you can see, this refers to an **intersection** of two independent events

i.e. $P(M \cap X) = P(M) \times P(X) = a/(a+b) \times (a+b)/N =$ **a/N**

3. Conditional Probability: If any of the individual figures is used against the sub-totals
e.g. Probability of *taking cocaine X times* **GIVEN THAT** one is a *male*
As you can see, this is a partition of a set into two mutually exclusive events, and dealing with one of the events

i.e. $P(X \mid M) =$ **a/(a+c+e)**

Resulting relationship: We can see here that; the probability of being male AND taking cocaine X times (Joint probability) is equal to the probability of taking cocaine X times GIVEN THAT one is a male TIMES the probability of being male:

a/N = a/(a+c+e) X (a+c+e)/N

$P(M \cap X) = P(X \mid M) \times P(M)$

This relationship is called the **Multiplication Rule**

Multiplication Rule: Joint probability = Conditional probability X Marginal probability
$P(M \cap C) = P(C \mid M) \times P(M)$

Example 2.3.1: Times taking cocaine by gender

Times taking cocaine	Male (M)	Female (F)	
X	A	B	**(a+b)**
Y	C	D	**(c+d)**
Z	E	F	**(e+f)**
	(a+c+e)	**(b+d+f)**	**N**

1. Marginal Probability: If any of the sub-totals is used as a numerator, and the grand total N as the denominator.

e.g. Probability of taking cocaine Y times (regardless of characteristic) = **(c+d)/N**

2. Joint Probability: If any of the individual cells is used against the grand total

e.g. Probability of *taking cocaine Y times* **AND** being a *female* = a/N.
As you can see, this refers to an **intersection** of two independent events

i.e. $P(Y \cap F) = P(Y) \times P(F) = (c+d)/N \times d/(c+d) =$ **d/N**

3. Conditional Probability: If any of the individual figures is used against the sub-totals

e.g. Probability of *being a female* **GIVEN THAT** one *takes cocaine Y times*
As you can see, this is a partition of a set into two mutually exclusive events, and dealing with one of the events
i.e. $P(F \mid Y) =$ **d/(c+d)**

Resulting relationship: We can see here that; the probability of taking cocaine Y times and being a female is equal to the probability of being female GIVEN THAT one takes cocaine Y times, TIMES the probability of taking cocaine Y times:

d/N = d/(c+d) X (c+d)/N

$P(Y \cap F) = P(F \mid Y) \times P(Y)$

The Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

c. Baye's Theorem

Illustration with a screening test: Imagine a screening test, such that the 2X2 table is as follows:

	Status		
	Diseased	**Not diseased**	
Result			
Positive	a = TP	b = FP	**(a+b)**
Negative	c = FN	d = TN	**(c+d)**
Total	**a+c**	**b+d**	**N**

Sensitivity: = TP/TP+FN = a/ (a+c) = $P(T \mid D)$ (Probability of testing positive GIVEN THAT one has disease)

Specificity: = TN/TN+FP = d/ (b+d) = $P(T^C \mid D^C)$ (Probability of testing negative GIVEN THAT one has no disease – you can see that it is complementary to sensitivity)

PPV: = TP/TP+FP = a/ (a+b) = P (D | T) (Probability of having disease given that one has a positive test

NPV: = TN/FN+TN = d/ (c+d) = P (D^C | T^C)

Now remember: **JP = CP X MP**

Thus: CP = JP/MP

Therefore: P (D∩T) = P (T | D) X P (D) It is also true that: P (D∩T) = P (D | T) X (PT)

Thus: $P(D \mid T) = \frac{P(D \cap T)}{P(T)}$

According to Baye's Theorem;

$$\mathbf{P(D \mid T) = \frac{P(D \cap T)}{P(T)} = \frac{P(T \mid D) \ . \ P(D)}{P(T \mid D)\,P(D) + P(T \mid D^C)\,P(D^C)}}$$

As you can see on substituting from the above 2X2 table:

$$\frac{a/(a+c) \ X \ (a+c)/n}{a/(a+c) \ X \ (a+c)/n + b/(b+d) \ X \ (b+d)/n} = a/(a+b)$$

Total probability

Assuming A_1 …… A_K are mutually exclusive and exhaustive events,

Then:

$$P(B) = P(B \mid A_1).\,P(A_1) + P(B \mid A_2).\,P(A_2) + \ldots\ldots + P(B \mid A_K).\,P(A_K)$$

$$\sum_{i=1}^{k} P(B \mid A_i).\,P(A_i)$$

You will find the concept of probability useful in hypothesis testing and in several other applications in epidemiology and biostatistics. In the probability approach, all ratios of events (including fractions, rates and proportions may be viewed as probabilities of those events happening).

2.4 Unit 4: INFERENTIAL STATISTICS II – ESTIMATION AND HYPOTHESIS TESTING

2.4.1 Introduction to the Unit

In the previous unit, we looked at probability distributions as the basis for statistical inference. We looked at the Standard Normal Distribution as the basis for understanding other distributions. As we noted earlier, the area of inferential statistics has 3 components: Sampling Distributions, Estimation and Hypothesis testing. In this unit, we shall look at the methods of estimation and hypothesis testing.

2.4.2 Unit Outline

The following topics will be covered:

1. Estimation
2. Hypothesis testing
3. Tests of Association for Contingency Tables
4. Sample Size Determination

2.4.3 Instructional goal

The student should be able to use point and interval estimates to test hypotheses about epidemiological associations so as to generalize from samples to populations.

2.4.4 Unit Objectives

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

1. Distinguish between point and interval estimation.
2. Know Type I and Type II error.
3. Formulate null and alternative hypotheses for testing differences in means or proportions.
4. Formulate a decision rule for testing a hypothesis.
5. Apply a test statistic, critical value, and p-value approaches to test the null hypothesis.

REQUIRED READINGS

Wayne W. Daniel (1998) Biostatistics: A Foundation for Analysis in the Health Sciences, 7th Edition (Published by John Wiley & Sons, Inc.)

Chapter 4: pages 83-99, 104-123.

Chapter 6: pages 161-166.

2.4.5 Time Frame

2 WEEKS

2.4.6 Content

Lesson 1: Estimation

Introduction: In real life, we cannot study an entire population. We use samples to infer on the characteristics of populations. A good estimate has two properties:

1. It is non-biased: Systematic error is minimal.
2. Variability from one sample to another. A good estimate is fairly stable from one sample to another. In normally distributed populations and larger samples, the mean is a better estimate than the median because it has a smaller standard error than that of the median.

The subject of estimation is concerned with methods by which population characteristics (parameters) are estimated from sample observations. Normally the population parameters are not known, and even where it is possible to be known, it may be expensive or infeasible to enumerate all the values of the population to determine the parameter in question. Statistical estimation procedures provide us with means of obtaining estimates of the population parameters with a desirable degree of precision. There are two different types of estimation:

a) Point estimation
b) Interval estimation → Confidence intervals

Definition of an estimator: An estimator is a statistical technique, rule or formula that we use to obtain the sample statistics.

Examples: a) the estimator of the population mean is given by: $\sum_{i=1}^{n} X_i \Big/ n$

b) the estimator of the population variance is given by

$$\text{Variance} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ (n-1)$$

$$\text{Standard Deviation} = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ (n-1)}$$

An estimator is a random variable because it is based on random observations (sample randomly selected). The result, or the value of an estimator, is what we call an estimate.

Lesson Topics: The following topics will be covered:

a. Point estimates
b. Interval estimates

Lesson Objectives:

By the end of this lesson, the student should be able to:

1. Evaluate and use point estimates in inferring from samples to populations.
2. Set up confidence intervals for means and proportions and interpret them.

REQUIRED READINGS

Wayne W Daniel (1998) Biostatistics: A Foundation for Analysis in the Health Sciences, 7th Edition (Published by John Wiley & Sons, Inc.)
Chapter 6; pages 150 - 199.

a. Point estimates

A point estimate gives a single value based on observations from a single sample. From it, we can determine the probability that a particular value will occur i.e., the proportion in the population that is likely to have a particular observation. It however gives no information about how close the value is to the population parameter. For instance, the sample mean $\bar{x}$ = **3** is a point estimate of the population mean, but it does not show us how close this is to the unknown true population mean.

Definition of point estimation: Point estimation is the statistical method (technique) which when applied gives a single value of the estimator, e.g. the mean.

If we have a point estimate of a sample parameter, we can determine the corresponding p-value i.e., the probability of its occurrence in a normal population. If the p-value is less than our set level of significance (e.g. 0.05 or 5%), we conclude that the sample parameter is different from the true population parameter and that the difference is statistically significant or could not have occurred by chance alone. Conversely, if its p-value is more than the set level, then the point estimate is not very different from the true population parameter. The use of point estimates and p-values in hypothesis testing will be discussed in detail in the next session.

b. Interval estimates

Introduction: It provides a range of values based on a sample of observations. It gives us information about the closeness of the sample estimate to the unknown population parameter. It is stated in terms of probability. Suppose a sample of 10 birth weights has been drawn. Our best estimate of the population mean μ would be the sample mean $\bar{x}$ = 116.9 kg. Although 116.9 kg is our best estimate of μ, we still are not certain that μ is 116.9 kg. Indeed, when another sample of 10 birth weights was drawn, a point estimate of 132.8 kg was obtained. Our point estimate would certainly have a different meaning if we were quite certain in some sense that the true population mean μ was within 1 kg of 116.9 rather than within 5 kg. We can construct intervals around the sample mean within which we are sure that to a certain probability extent that the population mean lies. We frequently then obtain an interval of plausible estimates of the mean as well as a best estimate of its precise value. Our interval estimates will hold only if the underlying distribution is normal; and only approximately, if the underlying distribution is not normal but can approximately be considered to be normal.

It is important to note at this point that the boundaries of the interval depend on the size and sample points chosen and will therefore vary from sample to sample.

Definition: Interval estimation is the method whereby the true value of the parameter is given between limits with a certain level of confidence. The confidence interval can also be looked at as a measure of precision.

Key Elements of Interval estimation: An interval estimate is a probability that the True population parameter falls somewhere within the interval. For instance, we can say that the observed sample estimate is 63 (the point estimate), and we are 95% confident that the true population parameter lies between 50 and 70 OR, among 100 repeated samples, this interval will include the true population parameter 95% of the times.

The Margin of error: The difference between the true population parameter and the sample parameter is called the margin of error or the precision of the estimate.

Factors affecting interval width:

1. Data dispersion: measured by the standard deviation.
2. Sample size
3. The level of confidence itself (90%, 95%, 99%)

Illustration:

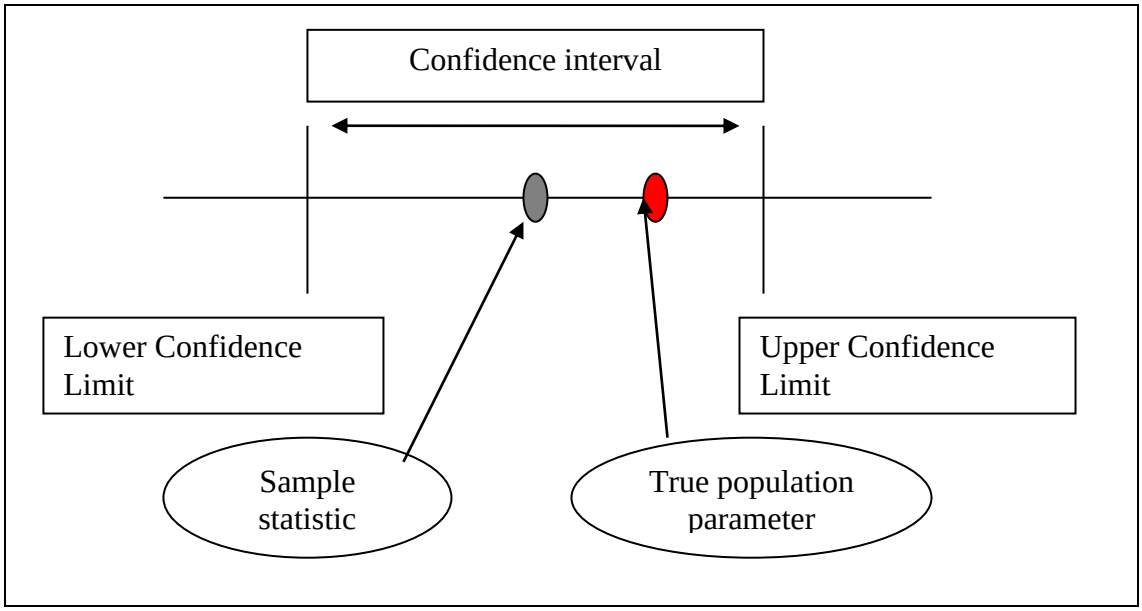
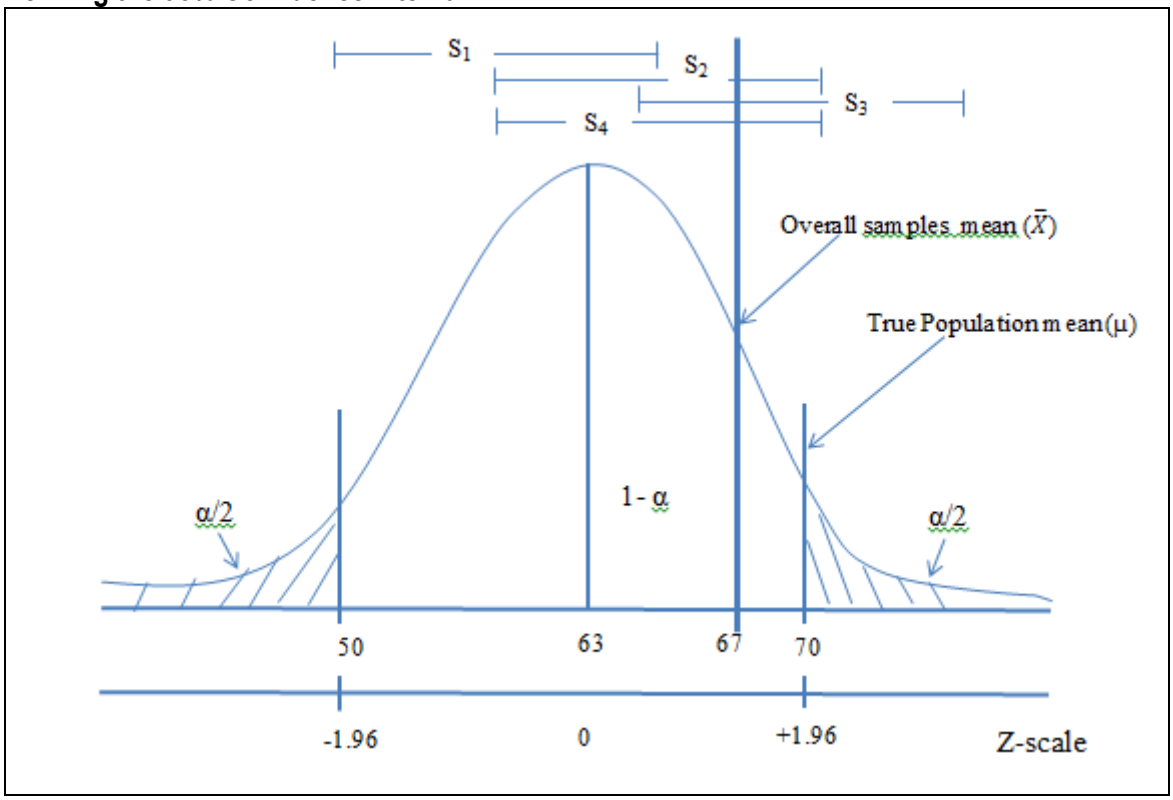


Illustration of the concept of the confidence interval: Since $\overline{X}$ is the mean of many samples (in this case 63), there is an interval on either side of $\overline{X}$ in which we are sure that the true population mean μ will be. (In this case, the arbitrary interval is 50 to 70; meaning that if we take 100 samples, the population-mean μ will lie between 50 and 70 in 95 of the samples. This interval is called the 95% Confidence Interval. This is further illustrated in the figure below:

Deriving the 95% Confidence interval



Constructing the 95% CI for a mean (Population standard deviation known) - we use the z-distribution: We now describe the rationale for the 95% CI;

We know that the z statistic is given by the formula:

$$z = \frac{(x_i - \bar{x})}{\sigma/\sqrt{n}}$$

If x_1 is the lower confidence interval limit (say 95%), the equivalent value for z after standardizing is – 1.96

If x_2 is the upper confidence interval limit (say 95%), the equivalent value for z after standardizing is + 1.96

Therefore: $-1.96 = \frac{(x_1 - \bar{x})}{\sigma/\sqrt{n}}$ and: $+1.96 = \frac{(x_2 - \bar{x})}{\sigma/\sqrt{n}}$

Cross multiplying: $-1.96 * \sigma/\sqrt{n} = (x_1 - \bar{x})$ and: $+1.96 * \sigma/\sqrt{n} = (x_2 - \bar{x})$

It follows then that: $x_1 = \bar{x} - 1.96 * \sigma/\sqrt{n}$ and: $x_2 = \bar{x} + 1.96 * \sigma/\sqrt{n}$

But: $\sigma/\sqrt{n}$ is the Standard Error, se

Thus: $x_1 = \bar{x} - 1.96 * se$ and: $x_2 = \bar{x} + 1.96 * se$

Therefore: Since x_1 and x_2 are the confidence interval range,

But we know that the population mean is between the two critical values: $x_1 > \mu > x_2$

Thus: $[\bar{x} - 1.96 * se < \mu < \bar{x} + 1.96 * se]$

It follows then that the confidence interval is given by the formula:

$$[\bar{x} \pm 1.96 * se] \text{ (for 95\% CI)}$$
or
$$[\bar{x} \pm z_{(\alpha/2)} * se] \text{ (for } (1-\alpha)\% \text{ CI)}$$

Assumptions: (When to use the Z-distribution in estimation:

i. That the population SD is known
ii. That the population is normally distributed
iii. That if not normal, it can be approximated by a normal distribution (n>30)

Example 2.3.9: The mean of a random sample of n=25 is $\bar{x}$ = 40. Set up a 95% Confidence interval estimate for the population mean if σ = 5.

$$[\bar{x} - z_{(\alpha/2)} * se < \mu < \bar{x} + z_{(\alpha/2)} * se]$$

Therefore: $40 - 1.96 * 5/\sqrt{25} > \mu > 40 + 1.96 * 5/\sqrt{25}$

38.04 < **μ** < 41.96

The 95% CI is therefore: **38.04** to **41.96**

Constructing the 95% CI for a mean (Population standard deviation not known) - we use the t-distribution: In the situation described before, we assumed that the population standard deviation is known and we used this to calculate the standard error of the mean. However, if the population standard deviation is not known, we use $s/\sqrt{n}$ instead of $\sigma/\sqrt{n}$ where *s is an estimate of* σ*:* Remember what we said earlier that the t-distribution is indeed a family of distributions which is affected by the ***degrees of freedom*. Therefore,** the value of the critical ratio varies with the sample size. At a sample size of 30 and above, it approaches the normal distribution curve and is **2.042** instead of 1.96. As the sample size increases to 150, this value in fact tends to 1.96.

It is therefore true that the population mean **μ** is between the two critical values:

$$[\bar{x} - t_{(\alpha),df(n-1)} * se < \mu < \bar{x} + t_{(\alpha),df(n-1)} * se]$$

Therefore, the formula for the confidence interval for a mean, using the t – distribution (population SD not known) is:

$[\bar{x} \pm 2.042, {}_{df(30)} * s/\sqrt{n}]$; (for 95% CI and a sample of size 31)

or

$[\bar{x} \pm t_{\alpha}, {}_{df(n-1)} * s/\sqrt{n}]$;(for $(1-\alpha)$ % CI at a sample size of n**)**

Understanding the concept of the degrees of freedom:
It is the number of observations that are free to vary after a sample statistic has been given
e.g.: If the sum of 3 numbers is 9
and X_1 is 4
X_2 is 2
Then X_3 is 3 (this cannot vary)

Put another way: The number of missing values which we can predict using a given series of other values, given other known parameters e.g. the mean or the sum.
e.g. if: 1+ 2 + X = 6 then the missing figure X is 3. If two figures were missing, we cannot predict them
if: 1+ 2 + X has a mean of 2 then the missing figure X is 3
We have the liberty to predict only one value. Therefore, the degrees of freedom in this case is n-1

Assumptions (when to use the t-distribution in estimation):

i. Population standard deviation not known
ii. Population should itself be normally distributed

Example 2.3.10: A random sample of n= 25 has $\bar{x}$ = 50 and the sample standard deviation (**s**) = 8. Set up a 95% confidence interval estimate for the population mean **μ**.

The formula is: $[\bar{x} - t_{(\alpha),df\,(n-1)} * s/\sqrt{n} < \mu < \bar{x} + t_{(\alpha),df\,(n-1)} * s/\sqrt{n}]$

Substituting: $[50 - 2.0639 * 8/\sqrt{25} < \mu < 50 + 2.0639 * 8/\sqrt{25}]$

[**46.69 < μ < 53.30**]

Therefore, the 95% CI for the estimate is: 46.69 to 53.30

Constructing the CI for a proportion - we use the z- distribution:

This involves categorical variables with two possible outcomes; success (possess a certain characteristic) and failure (do not possess the characteristic of interest). The fraction of the population in the success category is denoted by **P** whereas the proportion of the sample in the success category is denoted by ***p***.

$$P = \frac{X}{N} = \frac{\text{Number in the population with the characteristic}}{\text{Population size}}$$

From a sample, the population proportion is estimated by

$$p = \frac{x}{n} = \frac{\text{Number in the sample with the characteristic}}{\text{Sample size}}$$

When both np and n(1-p) are at least 5, p can be approximated by a normal distribution with mean $\mu_p = p$ and standard deviation $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$

In this case:

$$p - z_{(\alpha)} \cdot \sqrt{\frac{p.(1-p)}{n}} < P < p + z_{\alpha} \cdot \sqrt{\frac{p.(1-p)}{n}}$$

Therefore, the formula for the 95% CI is:

$$p \pm z_{(\alpha)} \cdot \sqrt{\frac{p.(1-p)}{n}}$$

Example 2.3.11: In a random sample of 200 males, 16 are overweight. Compute a 95% confidence interval estimate for the proportion of the population (P) that is overweight.

The sample proportion is therefore $\frac{16}{200}$ = 0.08

The formula is: $p - z_{(\alpha)} \cdot \sqrt{\frac{p.(1-p)}{n}} < P < p + z_{\alpha} \cdot \sqrt{\frac{p.(1-p)}{n}}$

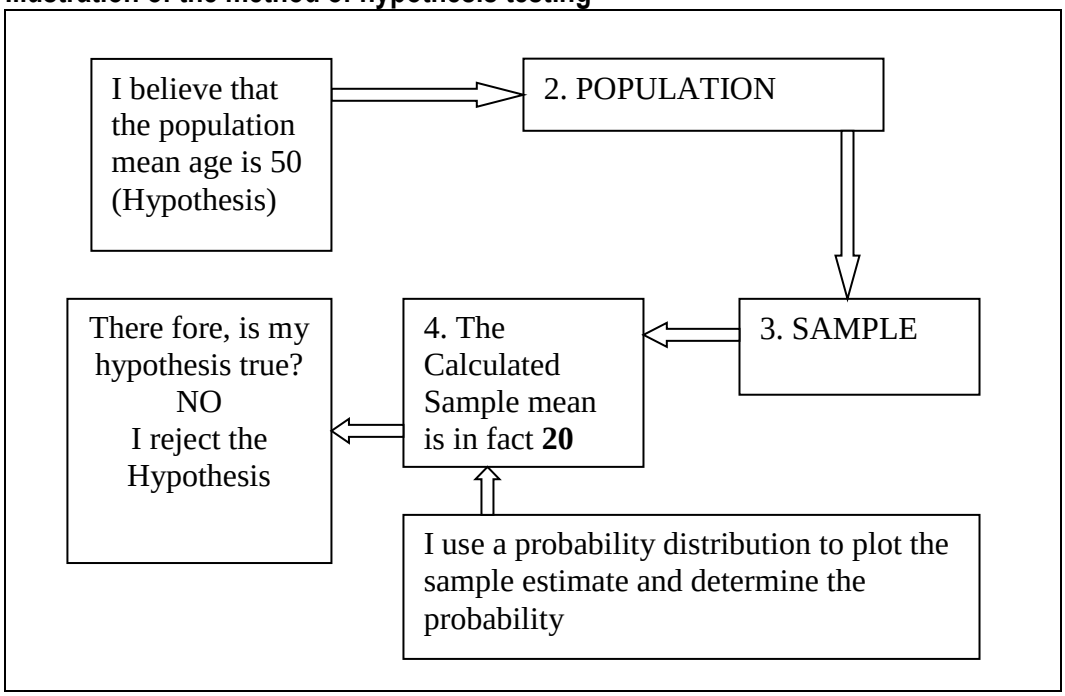
Substituting: **0.08 – 1.96.** $\sqrt{\frac{0.08\ .(1-0.08)}{200}}$ **< P< 0.08 + 1.96.** $\sqrt{\frac{0.08\ .(1-0.08)}{200}}$

Therefore: 0.042 < P < 0.118; the 95% CI is: 0.042 to 0.118

Lesson 2: Hypothesis testing

Introduction: To permit the generalization of sample parameters to the population, we use the point estimates or interval estimates to test hypotheses. From these estimates, we then use the corresponding probabilities to test whether the observed parameter can occur by chance alone (if the corresponding probability is less that 5% - assuming that a significance level of 95% is chosen), or indeed it approximates to the true parameter (If the corresponding probability is over 5% at the same level of significance) .To test these trends, we make use of hypothesis testing; the steps involved in hypothesis testing will be discussed in this lesson.

Illustration of the method of hypothesis testing



Lesson Topics: The following topics will be covered:-

The concept of hypothesis testing
The Z-test – population standard deviation known
Hypothesis testing scenarios – examples

Lesson Objectives:

By the end of this lesson, the student should be able to:

1. Defend the rationale for testing hypotheses as a tool used in statistical inference
2. Examine statements to discriminate between null and alternate hypotheses
3. Outline the steps involved in statistical hypothesis testing
4. Choose appropriate sampling distributions to test hypotheses about population characteristics, interpret the findings and draw conclusions on populations based on the sample

REQUIRED READINGS:

Wayne W Daniel (1998) Biostatistics: A Foundation for Analysis in the Health Sciences, 7th Edition (Published by John Wiley & Sons, Inc.)
Chapter 7, pages204 – 289.

a. Background information – the concept of hypothesis testing

What is a hypothesis? It is a belief about a particular Population Parameter. It must be stated before the analysis.

Definition: A hypothesis is a 'statement', 'idea' or an 'allegation' made about the true parameters of the population. When such a statement or idea is made, we may be interested in testing it to find out whether is true or not.

What is tested? We test H_0 (the null hypothesis). It has an equality sign, either = or ≤ or ≥
What is the opposite? It is H_A (alternate hypothesis). It has an inequality sign ≠ or < or >
How do you distinguish the two hypotheses? This depends on the question posed

Example 2.3.12: Test the hypothesis that the population mean is not 45.

Stated statistically: $\mu \neq 45$

In this case, the above cannot be a null hypothesis because it has an inequality sign. It is an alternative hypothesis.

The null hypothesis therefore is the opposite situation, but with an equal sign: μ = 45; i.e. the population mean is equal to 45

Example 2.3.13: The average amount of money spent in the bookstore per day is greater than 75 Shillings.

Stated statistically: $\mu > 75$ because the way it is stated, the expenditure is always greater than 75 (not inclusive of 75). This is an alternative hypothesis because it has no equality sign.

The Null hypothesis is therefore the opposite situation but with an equality sign: $\mu \leq 75$

Exercise 2.3.2: Identify the Hypotheses in these statements:

1. Is the population average length of stay for chronic disease patients 30 days?
2. Is the population average length of stay for chronic diseases different from 30 days?
3. Is the average number of heart beats **per minute less than or equal** to 85?
4. Is the average amount spent in the bookstore less than 95US Dollars?

The method of hypothesis testing: Suppose a paediatrician makes a statement that: "Mothers with low education level deliver babies whose birth weights are lower than normal". To test this hypothesis, a list of birth weights from consecutive, full-term, live-born deliveries from the maternity ward of Mulago hospital are obtained. In addition, information about the mothers highest level of education attained is collected, and the mothers grouped into 'low' and 'high' education level. The mean birth weight from the first 100 mothers categorized as low-education mothers is found to be 3.26 kg with standard deviation of 0.68 kg. Supposing we know from nationwide data that the nationwide mean birth weight is 3.4 kg and standard deviation of 0.71 kg. Can we say that the mean birth weight of low-education mothers is lower than average?

To test the paediatrician's hypothesis, we proceed by forming a null hypothesis H_o and an alternative hypothesis H_A. Thus for the above problem we can say that

H_o: The birth weight of babies from low-education mothers is not different from the national average.

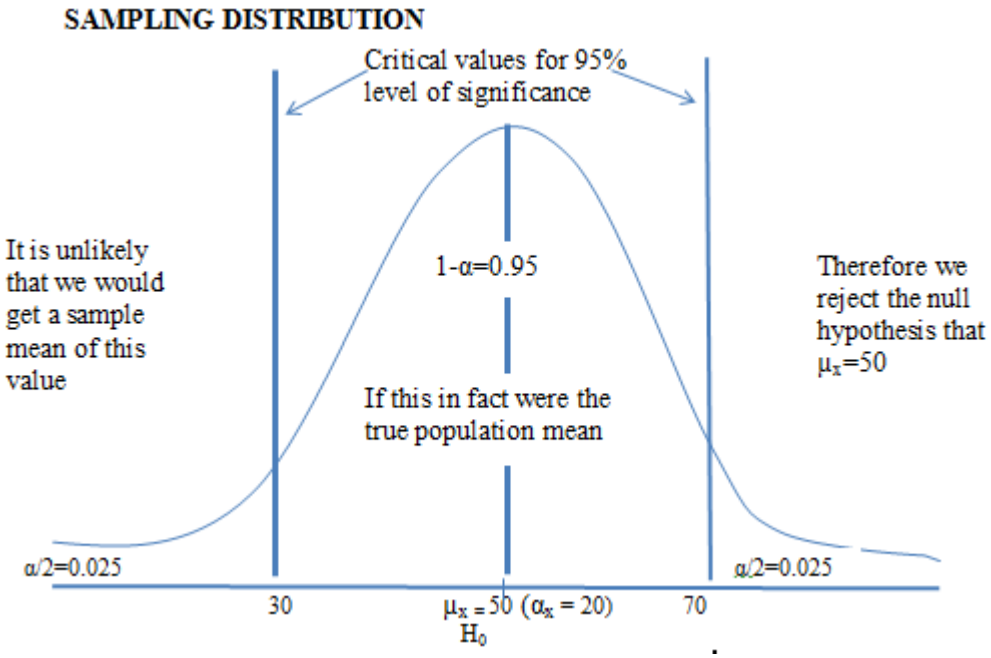
H_A: The birth weight of babies from low-education mothers is less than the national average.
(The alternative hypothesis in some sense contradicts the null hypothesis)

We proceed by assuming that the 100 birth weights from the low-education mothers come from an underlying normal distribution with known mean μ =3.4 kg and known standard deviation σ = 0.71 kg. We construct a one-sided (left-sided) 95% CI, based on the low-education mothers birth weights sample data. If the interval contains the national average mean, 3.4 kg, then the low-education birth weights are not significantly lower than the national average. In this case, we would accept the null hypothesis, and conclude that there isn't enough evidence to reject the null hypothesis. Note that this does not necessarily mean that the null hypothesis is correct.

Alternatively, if the 95% CI does not include the national average, then the low-education birth weights are significantly lower than the national average. In this case, we reject the null hypothesis and accept the alternative hypothesis as being correct.

b. Testing hypotheses using the z-test: Population standard deviation σ_X is known

Basic idea of hypothesis testing



The Level of significance: The level of significance is a probability. It defines the unlikely values of a sample statistic in the sampling distribution if the null hypothesis is true. It is called the rejection region of the sampling distribution and is designated as α**.** In most cases, we use 95% level of significance, therefore α = 5% or **0.05.** Other typical values of α may be 0.01 or 0.10. Please note that it refers to the area of the distribution that is beyond the critical values at that level of significance. This area is expressed as a proportion of the entire sampling distribution, as a percentage or a fraction. It is therefore a **probability**. The researcher selects the level of significance at the start.

The two-tailed test: In a two-tailed test, the level of significance is shared on either side of the sampling distribution, beyond the two critical values. The area beyond is therefore half of the selected **α** – value i.e. $(\alpha/2)$**.** We therefore have to bear this in mind if we are using **p-values** instead of the critical z-values, when checking the statistical tables for corresponding p-values – two tailed z-tables are straightforward because they directly give us a p-value based on both tails. On the other hand, when using one-tailed tables, we have to multiply the p-value by two, if we are testing a two-sided hypothesis. How do we know the kind of hypothesis? A two-sided null hypothesis has an equality sign e.g.:

H_0 : The population mean is equal to 50; therefore, H_A is that it is not 50; if it is not 50, it could be either way, either less or more – the probability is therefore two sided

Testing the Hypotheses: We use the following steps:

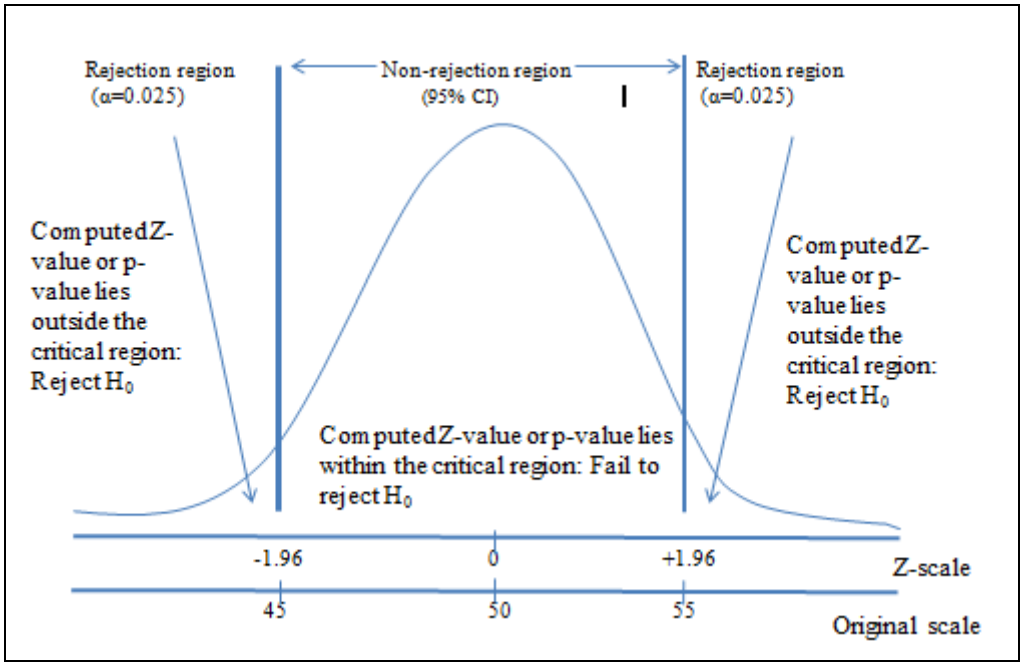
1. We convert the sample statistic (sample mean) to a standardized z variable, using the formula described before:

   i.e.: $z = \frac{x_i - \bar{x}}{\sqrt{\sigma^2}}$ or $z = \frac{x_i - \bar{x}}{\sigma}$

2. We compare the standardized value obtained to the critical z-values (e.g. + or – 1.96 for a two tailed z-test at 95% significance).
3. If the computed z-statistic falls beyond the critical values (i.e. if it is less than – 1.96 e.g. – 2.87/ OR if it is more than + 1.96 e.g. + 2.87, **we reject** the null hypothesis, H_0.
4. If the computed z-statistic falls within the critical values (i.e. if it is more than – 1.96 e.g. – 0.87/ OR if it is less than+ 1.96 e.g. + 0.87, **we fail to reject** the null hypothesis, H_0.

5. We can also use **probability** of an observation occurring: We obtain the z-statistic and check for its p-value. If it is less than 0.025 (one sided tables) or 0.05 (two tailed tables), we reject the null hypothesis, and accept the alternative hypothesis. If it is more than 0.05 (e.g. 0.07), fail to reject the null hypothesis.

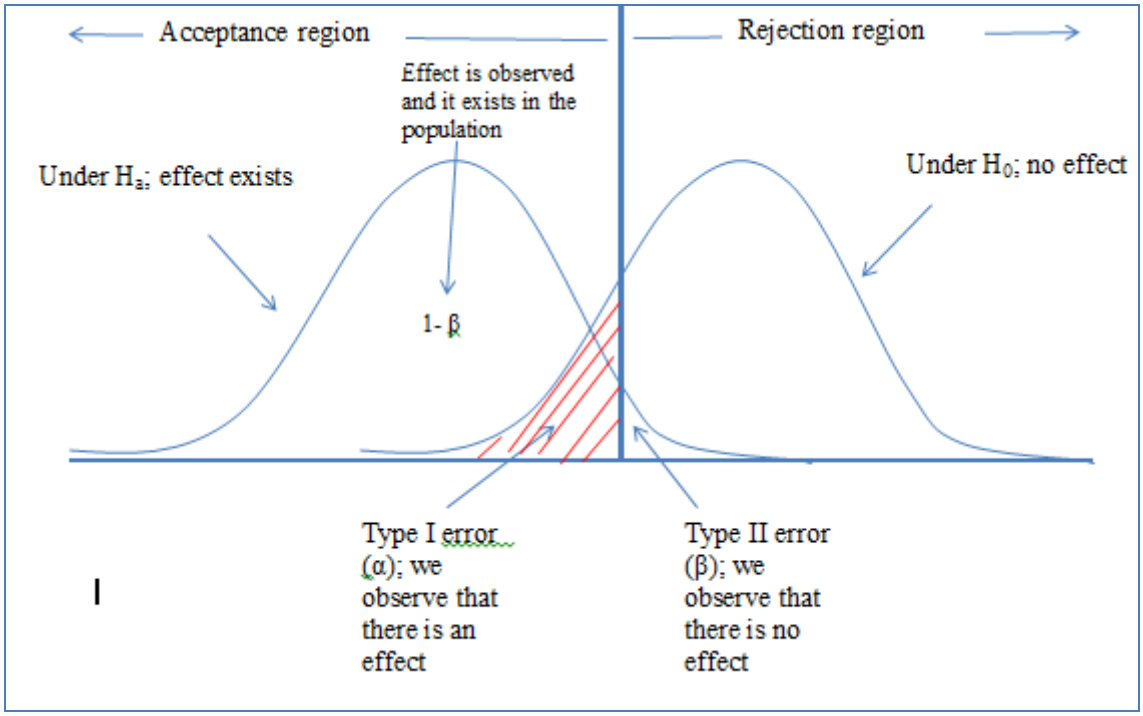
Illustration: Reject or accept null hypothesis at the 95% CI?



Errors in making a decision: When we test hypotheses and reject them or do not reject them we may make an error. There are two types of error:
1. Type I Error: This occurs when we reject a true null hypothesis; put another way, we say that there is an association when indeed there is none. This has more grave consequences. The probability of this is called the level of significance, denoted by α ; the higher the level of significance, the less this type of error. Therefore, in a type I error, we observe an EFFECT in the sample, when indeed there in NO EFFECT in the population.
2. Type II Error: This occurs when we do not reject a false null hypothesis; put another way, we say that there is an association when indeed there is none. The probability of this error is denoted β . Therefore, in a type II error, we observe NO EFFECT in the sample, when indeed there is an EFFECT in the population. The probability of its avoidance is $(1-\beta)$ and is referred to as the **power** of the study i.e. the probability of observing an EFFECT in a sample, when the EFFECT exists in the population.

These two types of error have an inverse relationship so that when we increase one, we reduce the other. The objective therefore is to strike a balance.

Illustration of the types of error:



Factors affecting β : There are 4 factors that will affect β **or** $(1-\beta)$**;** They are:

1. The true value of the population parameter: β increases when the difference between the hypothesized parameter and the true value decreases.
2. The significance level (α): β increases when α decreases
3. Population standard deviation σ: β increases when σ increases
4. Sample size n: β increases when n decreases

Summary of the Hypothesis testing steps: The following is the step wise approach to hypothesis testing that you will use in all cases, regardless of the distribution

Step 1: Evaluate the data
Step 2: Review the assumptions
Step 3: State the Null and Alternate Hypotheses
Step 4: Select the Test statistic
Step 5: Determine the distribution of this test statistic.
Step 6: Choose the level of significance, α and hence set up and check for the critical values or the critical probability; highlight then, the decision rule.
Step 7: Compute (standardize) the test statistic.
Step 8: Relate the computed value to the critical values or probability to make a statistical decision.
Step 9: Express the decision (conclusion) and interpret the result.
Step 10: Compute the corresponding p-value and report on it.

c. Examples of hypothesis testing approaches – different scenarios

1. One-population tests on numerical data: Using the z-test for means

Scenario one: Two - tailed Z test for mean (population standard deviation σ known)

a. Assumptions: Population is normally distributed

If not normal, it can be approximated by a normal distribution (n ≥ 30)

b. The null hypothesis has an equal (=) sign

c. Z-test statistic:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Example 2.3.14: Is the average pregnancy term 268 days? In a random sample of 25 women, the mean term $(\bar{x})$ was 272.5 days. Suppose from previous studies, σ was found to be 15 days. Test at the 0.05 level.

H_0: $\mu_X = 268$

H_A: $\mu_X \neq 268$

Test statistic: $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

$$z = \frac{272.5 - 268}{15 / \sqrt{25}} = 1.5$$

Critical values: At α = 0.05, the critical values are $\pm$ 1.96 (using two tailed z tables) [Remember: The given α in a two tailed z-test refers to a total of both sides of the distribution. In this case 0.025+ 0.025 = 0.05]

Decision: Since 1.5 is not beyond 1.96; we do not reject the null hypothesis

Conclusion: There is no evidence that an average pregnancy term is not 268 days

P-value: The corresponding p-value for a Z score of 1.5 **from the z-table** is 0.0668 for a **one tailed** table, corresponding to a cut off of $(\alpha / 2)$ = 0.025 for only this region. For a two tailed table, it is 0.1336; since this is greater than 0.05, it is in the non-rejection zone. We therefore do not reject the null hypothesis.

Scenario two: One - tailed Z test for mean (population standard deviation σ_X known)

a. Assumptions: Population is normally distributed

If not normal, it can be approximated by a normal distribution (n ≥ 30)

b. The null hypothesis has an equal/less/more than sign (≥ or ≤)

c. Z-test statistic

Z = $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

Example 2.3.15: Is it true that average pregnancy term does not exceed 268 days? In a random sample of 25 women, the mean term ($\bar{x}$) was 272.5 days. Suppose from previous studies, σ was found to be 15 days. Test at the 0.05 level.

H_0: $\mu_X \leq 268$

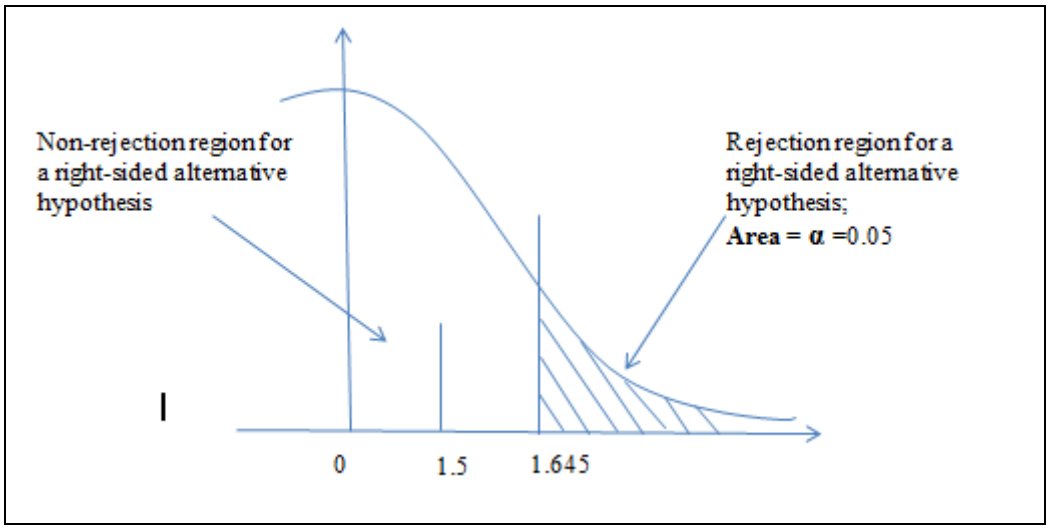
H_A: $\mu_X > 268$

Test statistic: $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

$$z = \frac{272.5 - 268}{15/\sqrt{25}} = 1.5$$

Critical values: At α = 0.05, the critical value is $\pm$ 1.645 (using one tailed z tables).It corresponds to one tail of the distribution. [If we use two tailed z tables, we first multiply the given α by two, then check the critical value. **NB:** In a one tailed test, it is the alternative hypothesis that points to the rejection zone i.e. if we state the null hypothesis above as: The average length of a pregnancy is at most 268 days (meaning that it is 268 or less and not more), then the alternative hypothesis is that the average length is greater than 268 days – when we are testing the hypothesis, the rejection zone is therefore in this direction (the tail to the right).

Illustration of directionality of rejection zone in a one tailed hypothesis

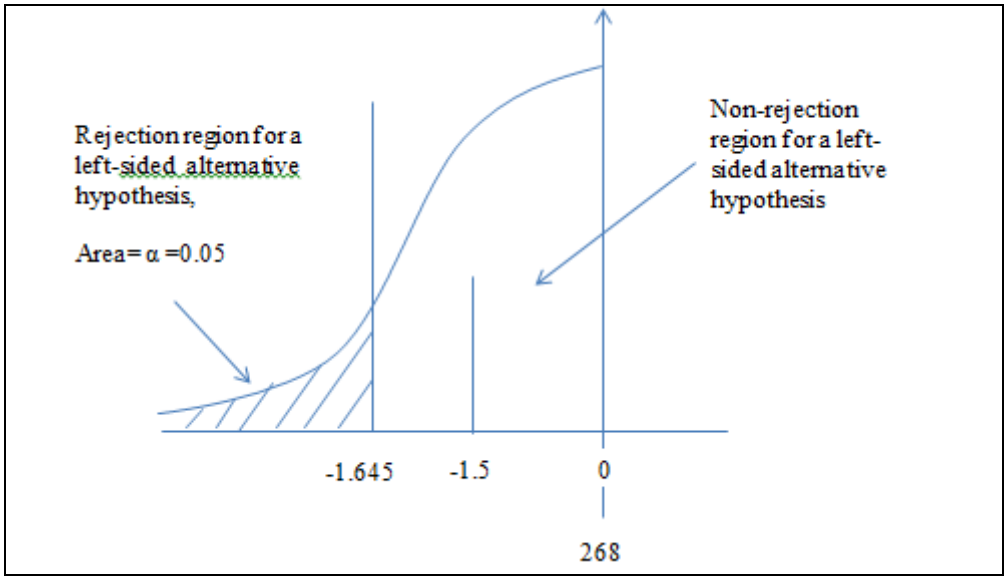


Decision: Since 1.5 is not beyond 1.645; we do not reject the null hypothesis.

Conclusion: There is no evidence from the sample that an average pregnancy term is more than 268 days; evidence points to the fact that it is at most 268 days.

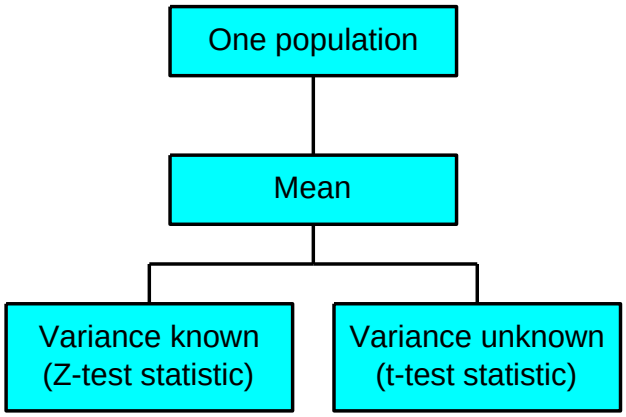
P-value: The corresponding p-value for a z score of 1.5 **from the Z-table** is 0.0668 for a **one tailed** table, corresponding to a cut off of α = 0.05 for only this region. For a two-tailed table, it is 0.1336, and therefore has to be divided by two before interpretation. Since the p-value of 0.06 this is greater than 0.05, it is in the non-rejection zone. We therefore fail to reject the null hypothesis.

If the null hypothesis above was instead as: The average length of a pregnancy is at least 268 days (meaning that it is 268 or more and not less), then the alternative hypothesis is that the average length is less than 268 days – when we are testing the hypothesis, the rejection zone is therefore in this direction (the tail to the left). This would be illustrated in the diagram below:



2. One sample tests on numerical data – the t-test

If we do not know the population standard deviation, or we do not take several samples, then we use the t-distribution to test hypotheses:



Scenario One: Two tailed t – test for mean (population standard deviation not known)

a. Assumptions: Population is normally distributed
If not normal, it is not largely skewed and can be approximated by a normal distribution ($n \geq 30$)

b. The null hypothesis has an equal (=) sign

c. t – test statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

d. Remember: The t-distribution is affected by the degrees of freedom

Example 2.3.16: Is the average pregnancy term 268 days? In a random sample of 36 women, the mean term ($\bar{x}$) was 272.5 days. The sample standard deviation is found to be 12 days. Test at the 0.05 level.

H_0: $\mu_X = 268$

H_A: $\mu_X \neq 268$

Test statistic: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{272.5 - 268}{12/\sqrt{36}} = 2.25$

The degrees of freedom are: $(n-1) = (36-1) = 35$

Critical values: At α = 0.05, and df 35, the critical values are $\pm$ 2.0301 (using two tailed z tables) [Remember: The given α in a two tailed z-test refers to a total of both sides of the distribution. In this case 0.025+ 0.025 = 0.05]

Decision: Since 2.25 is not beyond 1.96; we reject the null hypothesis

Conclusion: There is evidence that an average pregnancy term is not 268 days

P-value: The corresponding p-value for a t - score of 2.25 **from the two-tailed t-table** is between 0.02 and 0.05. Since this is less than 0.05, it is in the rejection zone. We therefore reject the null hypothesis.

Scenario Two: One tailed t – test for mean (population standard deviation not known)

a. Assumptions: - Population is normally distributed
- If not normal, it is not largely skewed and can be approximated by a normal distribution (n ≥ 30)

b. The null hypothesis has a less/more than/equal (≥) sign

c. t – test statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

d. Remember: The t-distribution is affected by the degrees of freedom

Example 2.3.17: Is the average capacity of batteries at least 140 ampere – hours? A random sample of 20 batteries had a mean capacity ($\bar{x}$) of 272.5 days. The sample standard deviation is found to be 2.66 days. Test the hypothesis at the $\alpha =$ 0.05 level.

H_0: $\mu \geq 140$

H_A: $\mu < 140$

Test statistic: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{138.47 - 140}{2.66/\sqrt{20}} = 2.57$

The degrees of freedom are: $(n-1)$ = 20-1 = 19

Critical values: At α= 0.05, and df 19, the critical values are – 1.7291 (using one tailed z tables) [Remember: The given α in a one tailed t-test refers to one side of the distribution, facing the direction of the alternative hypothesis.

Decision: Since – 2.25 is not beyond – 2.57 we reject the null hypothesis

Conclusion: There is evidence that the average capacity of batteries in the general population is not more than 140 ampere – hours. It is in fact less than 140 ampere – hours.

P-value: The corresponding p-value for a t - score of -2.57 **from the one tailed t-table** is even **less than 0.01**. Since this is less than 0.05, it is in the rejection zone. We therefore reject the null hypothesis.

3. One sample tests on categorical data – the z-test for proportions

Scenario One: Two - tailed Z test for mean (population standard deviation σ known)

a. Assumptions: Two categorical outcomes
If not normal, it can be approximated by a normal distribution (n ≥ 30)

b. The null hypothesis has an equal (=) sign

c. Z-test statistic:

$$z \equiv \frac{p - P}{\sqrt{\frac{p.(1-p)}{n}}}$$

Example 2.3.18: The present packaging system produces 10% defective cereal boxes. Using a new system, a random sample of 200 boxes had 11 defects. Does the new system produce fewer defects? Test at the $\alpha = 0.05$ level.

H_0: p ≥ 0.10

H_A: p < 0.10

Test statistic: $z \equiv \frac{p - P}{\sqrt{\frac{p.(1-p)}{n}}} = \frac{11/200 - 10/100}{\sqrt{\frac{0.10.(1-0.10)}{200}}}$

Direction of hypothesis: This is a one tailed hypothesis, and basing on the alternative hypothesis, it is left sided

Critical values: At α = 0.05, the critical values are – 1.645 (using one tailed z tables) [Remember: The given α in a one tailed z-test refers to the entire area on only one side of the distribution]

Decision: Since – 2.12 is beyond – 1.645, we reject the null hypothesis

Conclusion: We reject the null hypothesis and conclude that there is evidence that the new packaging system is less than 10% defective.

P-value: The corresponding p-value for a Z score of – 2.12 **from the z-table** is 0.0170 for a **one tailed** table. [In a two tailed table, it would be 0.0340 corresponding to a cut off of $(\alpha X 2)$ **= 0.10 or 90% CI;** If we used these tables, we would need to divide it by two to get the corresponding p-value for one side].

4. Comparing the means of two independent samples – the independent samples t-test for difference between means

Objective: We have two random samples, taken independently from two populations and our objective is to test for the difference between means.

Assumptions: The two populations are normally distributed with equal variances

Hypotheses: $H_0: \mu_1 = \mu_2$

$H_A: \mu_1 \neq \mu_2$

Sample characteristics	
Sample 1	**Sample 2**
n_1	n_2
Sample mean 1 = $\bar{x}_1$	Sample mean 2 = $\bar{x}_2$
Variance 1 = $s_1^2 = \dfrac{\sum(x_i - \bar{x}_1)}{n_1 - 1}$	Variance 1 = $s_2^2 = \dfrac{\sum(x_i - \bar{x}_2)}{n_{21} - 1}$
Pooled estimate of variance	
$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$	
But remember: The standard deviation for this relationship was: $\dfrac{\sigma}{\sqrt{n_1}} + \dfrac{\sigma}{\sqrt{n_2}}$ In this case, we use the sample standard deviation since we do not know the population one. Thus, $se = \dfrac{s_p}{\sqrt{n_1}} + \dfrac{s_p}{\sqrt{n_2}} = s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$	
Test statistic	
It has a t-distribution with degrees of freedom: $(n_1 + n_2 - 2)$ $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	

Let us now consider the example below:

Example 2.3.19: The table below shows the birth weights of children born to 14 heavy smokers **(group 1)** and to 15 non-smokers **(group 0)** sampled from live births at a large hospital.

Group 1 (Heavy smokers)	Group 0 (Non – smokers)
3.18	3.99
2.74	3.89
2.90	3.60
3.27	3.73
3.65	3.31
3.42	3.70
3.23	4.08
2.86	3.61
3.60	3.83

3.65 3.69 3.53 2.38 2.34	3.41 4.13 3.36 3.54 3.51 2.71

Solution: We substitute into the formulae

Sample characteristics	
Sample 1	**Sample 2**
N_1 = **14**	n_2 = **15**
Sample mean 1 = $\bar{x}_1$ = **3.1743**	Sample mean 2 = $\bar{x}_2$ = **3.6267**
Variance 1= $s_1^2 = \frac{\sum(x_i - \bar{x}_1) = 0.2145}{n_1 - 1}$	Variance1 = $s_2^2 = \frac{\sum(x_i - \bar{x}_2)}{n_{21} - 1} = 0.1285$
Pooled estimate of variance	
$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 0.1699$$ $$s_p = 0.4121$$	
But remember: The standard deviation for this relationship was: $\frac{\sigma}{\sqrt{n_1}} + \frac{\sigma}{\sqrt{n_2}}$ In this case, we use the sample standard deviation since we do not know the population one. Thus, $se = \frac{s_p}{\sqrt{n_1}} + \frac{s_p}{\sqrt{n_2}} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0.4121 * \sqrt{\frac{1}{14} + \frac{1}{15}}$	
Test statistic	
It has a t-distribution with degrees of freedom: $(n_1 + n_2 - 2)$ $t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{3.1743 - 3.6267}{0.4121\left(\sqrt{\frac{1}{14} + \frac{1}{15}}\right)} = -2.95$	

4. Comparing the means of two non-independent samples – the paired samples t-test for difference between means

Oftentimes, true differences do not exist between two populations with respect to the variable of interest, but the presence of extraneous sources of variation may cause rejection of the null hypothesis

of no difference. On the other hand, true differences also may be masked by the presence of extraneous factors. Paired comparisons would be employed to remedy such situations.
The objective in paired comparisons tests is to eliminate a maximum number of sources of extraneous variation by making the pairs similar with respect to as many variables as possible.

Objective: We have two random samples, taken from the same population and the objective is to test if there is a difference between the two means.

Assumptions: Unlike the independent samples t-test, analysis in the paired samples t-test is performed on d_i , the difference between the paired observations, as the variable of interest.

The differences d_i are normally distributed.

There is no need to worry about equality of variance since the variable is the difference between readings in the same individual, or matched individuals, and hence, only one variable is involved.

Hypotheses: $\mathbf{H_0: \mu_d = 0}$

$\mathbf{H_A: \mu_d \neq 0}$

Test statistic: $t = \dfrac{\bar{d} - \mu_{d_0}}{S_{\bar{d}}}$

$\bar{d}$ = sample mean difference

μ_{d_0} = hypothesized population mean difference

$S_{\bar{d}} = S_{\bar{d}} / \sqrt{n}$

Let us now consider the example below:

Example 2.3.19: The table below shows the test results of ten students prior to an intervention (pre-test) and the post intervention results (post-test). We would like to know if the intervention had asignificant effect on the sexual reproductive health knowledge levels of the adolescents. (Taken from the Iganga adolescent Reproductive Health Knowledge data set.)

ID	Pre-test result	Post-test result	d_i = (post-test – pre-test)	$(d_i - \bar{d})^2$
1	2	39	37	19.36
2	13	57	44	6.76
3	15	63	48	43.56
4	8	51	43	2.56
5	17	55	38	11.56
6	10	63	53	134.56
7	17	48	31	108.16
8	9	37	28	179.56
9	23	66	43	2.56
10	7	56	49	57.76
		Total	**414**	**566.4**

Sample characteristics
n = 10
Sample mean difference = $\bar{d}$ = 41.4
Variance = $s_{\bar{d}}^2 = \frac{\sum(d_i - \bar{d})^2}{n-1}$ = $\frac{566.4}{9}$ = 62.93
Test statistic : It has a t-distribution with $(n-1)$ degrees of freedom $t = \frac{\bar{d} - \mu_{d_0}}{S_{\bar{d}}} = \frac{41.4 - 0}{\sqrt{62.93/9}} = 5.92$
Critical value: $t_{(0.05,9)}$ = 2.262
Decision rule: Since the critical value of 2.262 is less than the calculated value of 5.92, we reject the null hypothesis.
Conclusion: There is a significant difference between the pre—test and the post-test results.

2.5 Unit 5: INFERENTIAL STATISTICS III: OTHER CONCEPTS

Lesson 1: Tests of Association for Categorical Outcomes

Introduction: In the previous lesson, we considered hypothesis testing methods for different scenarios in which the parameters are either means or single proportions. However, we can have situations in which we are comparing a multi-categorical outcome (2 or more levels of the outcome) in two or more populations. Our interest is to determine if any of the populations differs from the others significantly in relation to the outcome. In such cases, we use the tests for association for multi-categorical comparisons. These include:

- The chi-square tests of the association
- Point estimates of the measures of association; using Odds Ratios and Relative Risk
- Interval estimates for Odds Ratios and Relative Risk

In this lesson, we shall examine the use of chi-squares, as well as the method for constructing confidence intervals around Odds Ratios and Relative Risk.

Lesson Topics:

a. The chi-squared test
b. Hypothesis testing steps using the Chi-squared distribution
c. Confidence intervals for the measures of association

Lesson Objectives:

The MPH student should be able to:

1. Describe the chi-square test and the situations in which it is applicable
2. Make a step by step description of seven steps in hypothesis testing using the chi-square test
3. Use appropriate formulae to construct confidence intervals around Odds Ratios
4. Use appropriate formulae to construct confidence intervals around Relative Risks
5. Comparison of two proportion for as a test of association between variables

We will start by defining some basic tables:

a. **One-way classification**: A table in which the cells contain frequency counts of one categorical variable.

For example: 2x1

Take cocaine	Frequency
Yes	A
No	B
Total	**(A+B)**

b. **Two-way classification Table:** A contingency table that cross classifies two categorical variables. It is a joint frequency distribution of two binary categorical variables.

We compare a binary outcome with a binary independent variable.

For example: A screening test, the 2X2 table.

		Disease Status		
		Diseased	**Not diseased**	**Total**
Test Result	Positive	a	b	**(a+b)**
	Negative	c	d	**(c+d)**
	Total	**a+c**	**b+d**	**n**

c. There are situations in which we cross classify a binary outcome variable with a predictor variable possessing several categories. We can then construct *k*x2 contingency tables (k=number of categories of the independent variable) to represent the information.

An example is given below:

Cocaine use by depression, a 3x2 table

Times taking cocaine	Depressed (D)	Not depressed (N)	Total
X (Never)	A	B	**(a+b)**
Y (Sometimes)	C	D	**(c+d)**
Z (Always)	E	F	**(e+f)**
Total	**(a+c+e)**	**(b+d+f)**	**N**

d. **R × C Table**: A two-way classification table having R rows and C columns.

a. **The Chi square analysis of contingency tables**

The chi-squared test

We want to test the hypothesis that at **least one** category of people differs from others in the binary outcome and that the difference is statistically significant; that the categories are not homogenous in the classifying attribute. In this case, we compute the **chi-square (x^2)**. The Chi-square can be computed based on the formula below

where O_i = i^{th} observed frequency, E_i = i^{th} expected frequency in the i^{th} cell of a table with *k* cells

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Expected cell counts are given by:

$$a = \frac{(a+b)x(a+c)}{n}$$

$$b = \frac{(a+b)x(b+d)}{n}$$

$$c = \frac{(c+d)x(a+c)}{n}$$

$$d = \frac{(c+d)x(b+d)}{n}$$

The simplest computation is that of the 2X2 contingency table; referring to the table above, we compute the chi-squared as follows:

$$\chi^2 = \frac{(ad - bc)^2 N}{(a+b)(c+d)(a+b)(b+d)}$$

The values of χ^2 follow a **chi-squared distribution** with 1 degree of freedom. That is to say;

Degrees of freedom (df) = (Columns – 1) x (Rows – 1) i.e. (2 – 1) x (2 – 1) = 1

Corresponding probabilities can be checked from tables of the x^2 (chi-square) distribution. Also, in a 2x2 table, the x^2 value is equivalent to the square of the "Z" value in the Normal distribution. For example x^2 with 1 df (3.84) is equivalent to the root of $z^2(1.96^2)$, for the normal distribution
Note: The chi-square test is valid (or most appropriate) when the expected frequencies in any given cell is at least 5. If any cell has a frequency of less than five, we then use the **Fisher's exact test** which will be explained later.

b. Hypothesis testing steps using the Chi-squared distribution
Step 1: Determine the Hypothesis

H_0: $P_1=P_2$

H_1: $P_1 \neq P_2$

Step 2: Determine the statistical test to conduct; in this case, the Chi-squared test is appropriate because we are comparing a nominal dependent variable with a categorical independent variable.
Step 3: Assume α = 0.05

Step 4: Draw a contingency table and fill in the results summary; the degrees of freedom are $(2-1)(2-1)$ for a 2x2 table.

Step 6: Calculate χ^2

Step 7: Decision rule: If the observed value of χ^2 (df_1) is greater than the critical value in the Chi-squared distribution, then we reject H_0.

c. Confidence intervals for tests of association (Odds Ratio and Relative Risk)
From 2X2 contingency tables, we usually compute the Relative Risk (RR) and Odds Ratios (OR). These can be taken as point estimates. We can then construct confidence intervals around the point estimates.

Confidence interval for the Relative Risk: To obtain this, we first obtain a transformed form using the natural logarithm. The standard error of the natural logarithm of the Relative Risk is given by:

$$se(InRR) = \sqrt{\frac{1}{a} - \frac{1}{(a+b)} + \frac{1}{c} - \frac{1}{(c+d)}}$$

($InRR$ Refers to natural logarithm of the Relative risk)

Therefore, the 95% CI for lnRR is:

L1: $In(RR) - 1.96 * se(InRR)$

L2: $In(RR) + 1.96 * se(InRR)$

The Confidence interval for the RR is therefore the antilog $e^{L1}; e^{L2}$

Confidence Interval for the Odds Ratio: The standard error for the natural logarithm of the Odds Ratio [ln (OR)] is:

$$se(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Therefore, the 95% CI for *ln OR* is:

L_1: $In(OR) - 1.96 * se(InOR)$

L_2: $In(OR) + 1.96 * se(InOR)$

($\ln OR$ refers to natural logarithm of the Odds ratio)

The Confidence interval for the OR is therefore the antilog ($e^{L1}; e^{L2}$)

Confidence Interval for the difference between two proportions:

1-Hypothesis testing

Ho: $P_1 - P_2 = 0$ vs. Ha: $P_1 - P_2 \neq 0$

From the sample, calculate:
The SE of the difference in proportions is based on the pooled number of successes, "pbar".

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Then the test statistic is;

$$Z = \frac{\hat{p}_1 + \hat{p}_2}{\sqrt{\left(\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}\right)}}$$

We compare the observed Z with the tabulated Z- in the normal distribution tables, and then draw conclusions if there is an association between variables

2-Confidence interval based approach

When using this approach to assess association between variables, we construct a confidence interval based on the observed proportions to obtain the standard error of the difference between proportions. Then the 100(1 - α) % confidence interval for difference between proportion, p1-p2 is

$$\hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

If this confidence interval contains a zero, then there is no statistically significant association between variables of interest.

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Example 2.3.23: A cross-sectional survey was conducted to assess the respiratory health status of male farmers in Saskatchewan. The table below describes the relationship between self-reported asthma and use of carbamate insecticides.

		Asthma		
		Yes	***No***	
Carbamate	***Used***	32 *(38.6)*	465 (*25.1*)	497
	Not Used	51 (*61.4*)	1391 *(74.9*)	1442
		83	1856	1939

The values in bracket are the expected numbers in the cells under the null hypothesis that there is association between carbamate and Asthma. For example, the probability of having Asthma is 83/1939, and probability of use of carbamate is 497/1939. If there is no association between carbamate and Asthma, then the probability of carbamate use and having Asthma (32/1939) should be equal to the product of the probability of carbamate use and having asthma i.e. 32/1939 should be equivalent to (83/1939)*(497/1939). Therefore the expected frequency or number of individuals in this cell should be (83/1939)*(497/1939) *1939(total sample size). Therefore, for each of these cells, these expected frequencies are calculated- then we can use the formula above to obtain the chi-square value which can be compared with the tabulated value in the chi-square distribution tables.

Questions: 1) Test the hypotheses of no association between self-reported asthma and use of carbamate insecticides.

2) Calculate the Relative Risk of asthma.

3) Calculate the Odds Ratio if asthma for carbamate and the 95% confidence limits for the Odds Ratio.

4) Calculate the estimated attributable risk of asthma related to the use of carbamate insecticides.

Results:

(1) x^2 = $\frac{(32x1391 - 465x51)^2}{497x1442x83x1856}$

x^2 = 7.596, $(df = 1)$ and p=0.006. We reject the hypothesis of no association and conclude that the use of carbamate insecticides is significantly related to self-reported asthma.

(2) Estimated Relative Risk (RR) = $\frac{32/497}{51/1442}$ = 1.822

(3) Estimated Odds Ratio (OR) = $\frac{32X1391}{465X51}$ = 1.878

(4) The 95% confidence limits for ln (OR) are given by

(0.630 – 1.96X 0.232; 0.630 + 1.96 X 0.232) = (0.175, 1.085)

Therefore, the 95% CI for the OR is ($e^{0.175}$, $e^{1.085}$) = (1.19, 2.96)

Difference in proportions

3-Comparing differences in two proportions *Hypothesis Testing*

Ho: p_1-p_2=0 vs. Ha: p_1-$p_2 \neq 0$

From the sample, calculate:

The test statistic:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

The SE of the difference in proportions is based on the pooled number of successes, "pbar"

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

3-Comparing differences in two proportions *Confidence interval*
100(1 - α) % confidence interval for p1-p2 is

$$\hat{p}_1 - \hat{p}_2 \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

The SE of the difference in proportions is ***NOT*** based on the pooled number of successes, "pbar"

Lesson 4: Sample Size Determination

Introduction: Determining the appropriate sample sizes is another important aspect of inferential statistics. We ought to determine a sample size that is adequate to detect the change we are interested in. In this lesson, we shall discuss the rationale for deriving the commonest formulae for sample size determination. However, only introductory information will be given. Details of sample size determination will be covered in Applied Biostatistics II.

Lesson Topics

a. *Estimating the mean of a continuous outcome*
b. *Estimating the proportion of a binary outcome*
c. *Sample size with power*

Lesson Objectives:

The MPH student should be able to:

1. Calculate correct sample sizes required for detection of outcomes where the parameter is a mean, using a given set of requirements
2. Calculate correct sample sizes required for detection of outcomes where the parameter is a proportion using a given set of requirements
3. Determine sample sizes required for studies in which power considerations are incorporated in the design

a. Sample size determination when estimating the mean of a continuous outcome

Remember that: $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

But remember: The sample mean is an estimate of the true population mean. The difference between the two is called the margin of error.

Thus: $(\bar{x} - \mu)$ = Margin of Error; Therefore: $z = \frac{Error}{\sigma/\sqrt{n}}$

Cross multiplying: $z = \frac{Error.\sqrt{n}}{\sigma}$

Cross multiplying again: $\frac{z\ .\sigma}{Error} = \sqrt{n}$

Squaring all terms to remove root: $n = \frac{z^2.\sigma^2}{Error^2}$

Example 2.3.20: What sample size is needed to be 90% confident of being correct within ± 5? A pilot study suggested that the standard deviation is 45.

$$n = \frac{z^2.\sigma^2}{Error^2}$$

$= \frac{1.645^2 \,.\, 45^2}{5^2} = 219.2 \equiv 220$

b. Sample size determination when the estimating the prevalence/proportion of a binary outcome

The formula is: $n = \frac{z^2.\sigma^2}{Error^2}$ But: $\sigma^2 = p(1-p)$

Therefore: $n = \frac{z^2.p(1-p)}{Error^2}$

Note: $p(1-p)$ is a maximum when p=1/2, and it is **0.25**

Example 2.3.21: A new vaccine is to be tested on the market. Find how large a sample should be drawn if we want to be 90% confident that the estimate will not be in error of true proportion of success by more than 0.1. Earlier studies with a similar antigen showed a success of 50%.

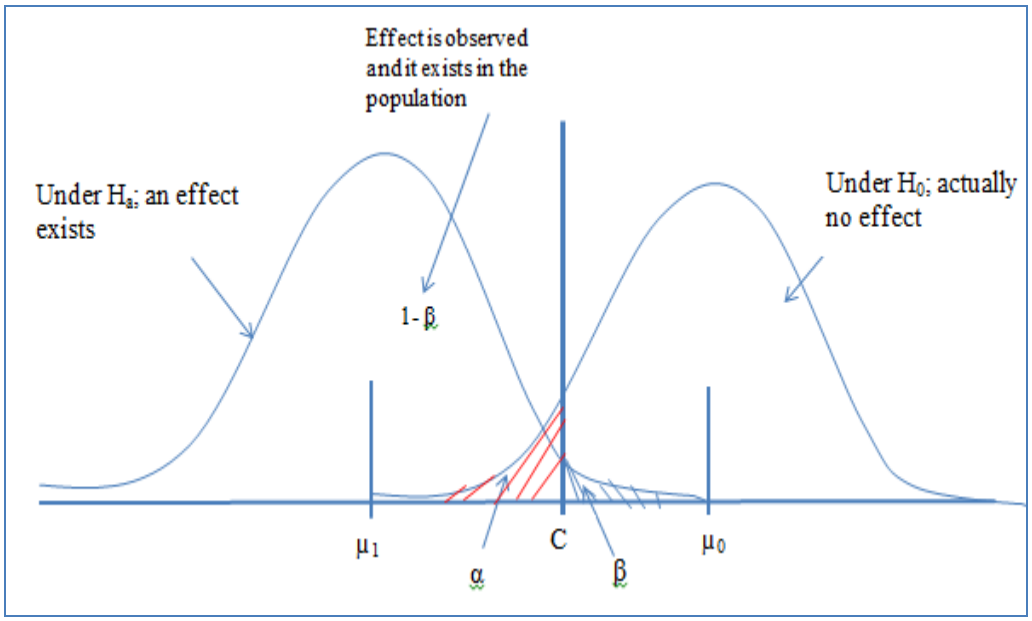
$$n = \frac{z^2.p(1-p)}{Error^2}$$

$= \frac{(1.645)^2 \,.\, (0.25)}{(0.1)^2} = 41.12 \equiv 42$

c. Sample size with power

Remember the α and β error curves; the two functions work in inverse, such that when we increase one, we reduce the other. The critical cut off point is shared by both curves and we can construct a simultaneous equation from it:

Illustration:



Using the critical Z value for power (β), we can construct an equation for C:

$$c = \mu_1 + z_1.\sigma/\sqrt{n} \quad \text{..... (1)}$$

But also
Using the critical Z value for alpha, we can construct an equation for C:

$$c = \mu_0 + z_0.\sigma/\sqrt{n} \quad \text{..... (2)}$$

Therefore, since the C is shared; $[\mu_1 + z_1.\sigma/\sqrt{n} = c = \mu_0 + z_0.\sigma/\sqrt{n}]$;

This implies that: $n = \dfrac{(z_0 + z_1)^2.\sigma^2}{Error^2}$

Where *Error* = $(\mu_0 - \mu_1)$

NB: At a power of 50%, the corresponding critical Z value is 0, implying that this has no effect on the sample size.

Example 2.3.22: What sample size is needed to be 90% confident of being correct within +/-5. A pilot study suggested that the standard deviation of 45; we want to be able to have a power of 80% to defect an effect if indeed it exists.

$$n = \frac{(z_0 + z_1)^2.\sigma^2}{Error^2}$$

$$= \frac{(1.645 + 0.84)^2 \,.\, (45)^2}{(5)^2} = 500.2 \equiv 501$$

Remember the factors affecting β ! There are 4 factors that will affect β **.** They include

1. The true value of the population parameter: β increases when the difference between the hypothesized parameter and the true value decreases.
2. The significance level (α): β increases when α decreases
3. Population standard deviation σ: β increases when σ increases
4. Sample size n: β increases when n decreases

Lesson 5: Sampling Procedures

In your additional resources folder, search for two articles on sampling methods:

- A PDF article titled: Overview of Sampling Procedures by the Fairfax County Department of Systems Management for Human Services, April 2003.
- A presentation by Dunstan Bagenda on 'Sampling'

Try and understand the following:

- Why we sample
- Probability sampling methods
    - Simple Random Sampling
    - Stratified random sampling
    - Cluster sampling
    - Systematic sampling
    - Multistage sampling
- Non probability sampling methods
    - Purposive sampling
    - Quota sampling
    - Convenience sampling
    - Judgemental sampling

2.4.7 Extension Activities

Extension Activity 1: Discussion Forum Question
1. *Statistical Methods make use of several sampling distributions in testing hypotheses about population characteristics using sample information. Select one sampling distribution and briefly describe its characteristics, parameters and when it is used* 2. *There are diverse hypothesis testing scenarios that depend on the directionality of the hypothesis to be tested as well as whether certain population parameters are known. Select one scenario and illustrate, with a numerical example, its salient features*

Extension Activity 2: Self- Assessment Quiz

QUIZ 2.3.1

Standardizing the sampling distribution of means and determining probability

Q1.The average age at which a woman has her first birth in a certain city is 26 years, with a standard deviation of 5.6. A random sample of 45 women is selected. What is the probability that the average age at which these women had their first child was between 24 and 25 years?

Standardizing the sampling distribution of the difference between means and determining probability

Q2. A comparison study of female and male patients also reported measurements of Inspiratory Vital Capacity (IVC). The results were as follows:

Men: $\mu_1 = 2.8$ $\sigma_1 = 0.8$

Women: $\mu_2 = 1.9$ $\sigma_2 = 0.5$

In a random sample of 16 men and 25 women, what is the probability that the difference between sample means will be more than 1?

Standardizing the sampling distribution of proportions and determining probability

Q3. Suppose it is known that in a certain population, 55% of people believe that cancer is curable. If a random sample of 200 is drawn, what is the probability that the sample proportion of people who believe that cancer is curable will be less than 0.45?

Interval estimates

Q4. The mean time required for 25 ambulances to reach their destinations was 7 minutes. Set up a 95% Confidence interval estimate for the population mean if $\sigma = 3$.

Q5. In a sample of 20 people, the mean diastolic blood pressure is 68.7 and the sample standard deviation is 13.02. Set up an interval estimate in which we are 95% confident that the true population parameter lies

Q6. In a random sample of 1000 people with angina pectoris, 80 are overweight. Compute a 95% confidence interval estimate for the proportion of the general population (P) that is overweight.

Q7. Suppose a woman wishes to estimate her exact day of ovulation for contraceptive purposes. A theory exists that at the time of ovulation the body temperature raises by an amount from 0.5^0 F to 1.0^0 F. Thus, changes in body temperature can be used to roughly estimate the day of ovulation. To use this method, a good estimate of basal body temperature during a period when ovulation is definitely not occurring is needed. Suppose that for this purpose a woman measures her body temperature on awakening for 100 days after menstruation and obtains a mean reading of 97.2^0 and standard deviation of 0.2^0. Construct a 95% CI for the underlying mean basal body temperature.

Q8. Re-construct the 95% CI in the above example assuming that the number of readings (n) was only 10 instead of 100; what is your comment about the confidence interval?

Q9. Re-construct the 95% CI in the above example assuming that the standard deviation of basal temperature is 0.4 rather than 0.2; and that the number of readings (n) was only 10 instead of 100. What is your comment about the confidence interval?

Hypothesis testing

Q10. Is the average pregnancy term 268 days? In a random sample of 25 women, the mean term ($\bar{x}$) was 272.5 days. Suppose from previous studies, **σ** was found to be 15 days. Test at the 0.05 level.

Q11. Is it true that the average period of stock-out of all three indicator essential drugs does not exceed 268 days? In a random sample of 25 health units, the mean stock-out time $(\bar{x})$ was 272.5 days. Suppose from previous studies, **σ** was found to be 15 days. Test at the 0.05 level.

Q.12 Is the average pregnancy term 268 days? In a random sample of 36 women, the mean term $(\bar{x})$ was 272.5 days. The sample standard deviation is found to be 12 days. Test at the 0.05 level.

Q13. Is the average capacity of theatre lamp inverters at least 140 ampere – hours? A random sample of 20 inverters had a mean capacity $(\bar{x})$ of 272.5 days. The sample standard deviation is found to be 2.66. Test the hypothesis at the 0.05 level.

Q14.The present drug packaging system produces 10% defective essential kit boxes. Using a new system, a random sample of 200 boxes had 11 defects. Does the new system produce fewer defects? Test at the 0.05 level.

Q15. The data in the table below was taken from a case-control study on the relationship of cataract with diabetes.

DIABETES	**Cataract cases**	**Fracture patients (Controls)**
Present	55	84
Absent	552	1927

(Taken from Statistical Methods in Epidemiology, page 57)

(1) Test the hypothesis that the probability of diabetes is the same in both cases and controls
(2) Calculate the Odds Ratio of Cataract and 95% confidence limits for the Odds Ratio for the diabetic subjects

2.6 Unit 6: ANALYSIS OF VARIANCE

2.6.1 Introduction to the Unit

Background: In Applied Biostatistics and Informatics I, you were introduced to the t-test, a bivariate technique. In fact, the t-test was valuable in comparing means from two samples. If the samples were independent, then we performed and ***independent samples t-test***. If the samples were dependent (before and after) we performed a ***paired t-test***. All in all, the purpose of the t-test was to compare the difference between the means of two samples, and draw statistical inference if indeed their apparent difference is statistically significant. An example is given below:

Illustration: In a Field trial of Pedagogical approaches for Adolescent RH Education in Secondary Schools, the mean class scores for a Lecture Instructed arm were compared with the mean class scores for a participatory instructed arm. The following results were obtained:-

Class	Participatory	Lecture
Senior 1	49.7	42.2
Senior 2	48.9	35.8
Senior 3	50.5	46.7
Senior 4	45.4	39.7
Senior 5	52.8	35.5
Senior 6	42.1	21.1
Mean	**48.2**	**37.0**

The difference between means in this case was 11.2 marks, with a 95% Confidence Interval of 9 to 14 marks using the independent samples t-test. Students who underwent participatory instruction scored an average of 11 marks higher than those that underwent lecture style instruction and the difference was statistically significant.

The situation just described is relevant to a situation in which we have to compare means from only **two samples**. Very often in statistics and epidemiology, we have to compare the means of several independent samples (more than two). In such cases, a instead of using a t-test, we employ the techniques of Analysis of variance. ***Analysis of variance is therefore a relevant statistical test if we have a multi-categorical independent variable, and numerical dependent variables.***

2.6.2 Unit Outline

The following Sessions will be covered:

Session 1: Introduction to ANOVA

Session 2: One-way ANOVA (ANOVA for the Completely Randomized Design)

2.6.3 Instructional goal

The MPHO should be able to apply the method of ANOVA in statistical hypothesis testing where it is indicated

2.6.4 Unit Objectives

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

1. Define and describe the meaning and application of ANOVA
2. Set up and evaluate an ANOVA for a Completely Randomized Design
3. Set up and evaluate an ANOVA for a Randomized Block Design

2.6.5 Time Frame

1 WEEK

2.6.6 Content

Session 1: Introduction to ANOVA

Introduction: In this Session, we shall familiarise ourselves with an important Bi-variate and Multivariate technique used to test whether there is a difference between means in a multi-categorical array of independent samples (More than two categories). Before we go on to discuss the methods, we shall define and describe what the method of ANOVA entails.

Session Topics: The following topics will be covered:

a. What is ANOVA?
b. Types of Experimental Study Designs
c. The Variables in ANOVA
d. Testing Hypotheses using the ANOVA Procedure
e. Calculation of the test statistic in ANOVA

Session Objectives:

By the end of this Session, the MPHO should be able to:

1. Define ANOVA
2. Differentiate between the Completely Randomised Design and Randomised Blocks
3. Distinguish the different types of variables in ANOVA
4. Outline the key steps involved in testing hypotheses using ANOVA

a. What is ANOVA?

Analysis of Variance, commonly referred to as ANOVA (uh-nove-uh), is analogous to a between groups t-test that is used with two groups. It is a more general test, though, that allows one to compare several groups at once, not just two. Instead of using the t-statistic as in the t-test, we use an F statistic. Why F?

The letter F is a tribute to the inventer of ANOVA, *Sir Ronald Fisher*, knighted for his accomplishments in statistics. Fisher was a statistician who studied agricultural genetics among other things. It turns out that the F-test (or ANOVA) with two groups is analogous to the t-test. You'll get the same result with either. But the ANOVA test is more general because it can be used in more complex studies that compare more than two groups. In Applied Biostatistics I, the t-statistic was described as a type of ratio- a ratio between the group difference and sampling variability (i.e. the standard error)

$$t = \frac{Difference\,Between\,Groups}{Sampling\,Variability}$$

In fact, the t – Statistic is given by: -

$$t = \frac{\bar{x} - \mu_x}{\sigma^{\bar{x}}} = \frac{\bar{x} - \mu_x}{\sigma_S / \sqrt{n}}$$

It was mentioned that the standard error is really based on the standard deviation, a measure of variability within the sample. ANOVA is really based on the same idea, but Fisher conceptualized it slightly differently. He thought of it as a ratio of two types of variances, the variance between group means and overall variance in the sample.

$$F = \frac{Variance\ Between\ Groups}{Overall\ Variance}$$

In the between groups t-test, we examined the difference between two means $(\bar{y}_1 - \bar{y}_2)$. Another way to think of the difference between two means is as a type of variation among the means. Here we have just two means, but a difference is the same thing as a variation. (Remember the calculation of the variance of a group of numbers involves subtracting y from $\bar{y}$?). If there are several groups, their group means may differ or vary. The overall variance is just the variation of scores in the sample. So, we could restate the F-test this way:

$$F = \frac{Variance\ Among\ the\ Group\ Means}{Difference\ Among\ the\ Scores}$$

If the differences between the group means are large relative to the amount of variability in the scores, the group differences are probably significant. If, however, there is a lot of variability in the scores, then the difference between (or among) the group means will not seem so large, a non-significant difference.

In general, ANOVA is defined as a technique in which the total variation present in a set of data is partitioned into two or more components. Associated with each of these components is a specific source of variation, so that the in the analysis, it is possible to ascertain the magnitude of the contribution of each of these sources to the total variation and draw conclusions on it. It has wide application in the analysis of experimental studies that involve measurement of continuous/numerical among more than two groups.

Use of ANOVA: ANOVA is used for two different purposes: To estimate and test hypotheses about population variances and to test and estimate hypotheses about population means. In most cases, the latter is often used. However, as we see, our conclusions on the magnitude of the means are dependent on the magnitude of the observed variances.

b. Types of Experimental Study Designs

ANOVA finds widest application in experimental study designs. There are three types of designs: -

- The Completely Randomised Design
- The Randomised Block Design
- The Factorial Design (Latin Square)

In this course, we shall concentrate on these two.

The Completely Randomised Design (CRD): Treatments are applied to each of the study units individually, and each study individual is randomly allocated to an intervention category.

The Randomised Block Design (RBD): In this design, the study participants are organised in blocks within each study category. The blocks are either strata or clusters, discrete in nature.

c. The Variables in ANOVA

We have three types of variables in ANOVA:

- The Independent variable
- The dependent variable
- The extraneous variable

The Independent variable – It is also called the Treatment variable. It refers to the intervention category to which an individual has been allocated. It is therefore a categorical variable (a nominal category) e.g. Intervention Category.

The Dependent variable – It is the outcome or response variable and is often a continuous variable, comprising of a set of discrete numerical measurements from the study units.

The Extraneous variable – They are variables that may have an effect on the outcome variable but are not the focus of this study. They therefore lead to a measurable level of "experimental error" in each measurement.

Example: In a Field trial of Pedagogical approaches for Adolescent RH Education in Secondary Schools, the mean class scores for a Lecture Instructed arm and a Participatory arm were compared with the mean class scores for a Control Arm. The following results were obtained:-

	Independent Variable (Treatment or Intervention)		
Dependent Variable	**Participatory**	**Lecture**	**Control**
Class			
Senior 1	49.7	42.2	2.3
Senior 2	48.9	35.8	4.8
Senior 3	50.5	46.7	7.4
Senior 4	45.4	39.7	4.7
Senior 5	52.8	35.5	4.7
Senior 6	42.1	21.1	5.1
Mean	**48.2**	**37.0**	**4.9**

d. Testing Hypotheses using the ANOVA Procedure

In applied Biostatistics and Informatics I, we leant about the 9- step hypothesis testing procedure. In testing hypotheses using the F-test, we also use a nine step process as follows:

1. ***Description of the Data:*** We have to make an overall description of the data by displaying it, preferably in tabular form and examining it.
2. ***Assumptions:*** In order for ANOVA to be conducted as a parametric test, there are a set of assumptions that have to be met, concerning the data. Based on these assumptions, we then have to select a model
3. ***Hypotheses:*** We have to state the Null Hypothesis and the Alternative Hypothesis
4. ***The Test Statistic:*** We select the test statistic. In ANOVA, it is the ratio of the variances of the different sources of variation
5. ***Determine the distribution of the test statistic:*** For ANOVA, it is the F-Distribution.
6. ***Decision Rule:*** We have to decide on how the decision will be made; this is based on the level of significance selected.
7. ***Calculation of the test statistic:*** We then conduct arithmetic calculations for the different sources of variation that form the components of the test statistic. We summarise these components in the **ANOVA Table**. We then calculate test statistic.
8. ***Statistical Decision:*** We then make the statistical decision.
9. ***Conclusion:*** From the statistical decision, we express the conclusion in narrative terms
10. ***P-Value:*** We can determine the p-value and report on it.

The F- Distribution:

The F- Distribution is a skewed distribution of the ratio of variances from different sources of variation. The distribution is affected by Degrees of freedom.

ACTIVITY: Using statistical tables, describe the F-Distribution, its characteristics and values.

e. Calculation of the test statistic in ANOVA

You will recall that during applied Biostatistics and Informatics I, we learnt about calculation of different test statistics. The test statistics were indeed a form of standardisation of observed parameters, so that they can be compared on the same scale in a particular probability distribution. It is these test statistics that we then used for making statistical decisions, using the appropriate probability distribution and the "Critical Values" or "Levels of Significance".

The process in ANOVA is no different. We have to compute the test statistic, and use this in statistical decisions. The computation is dependent on whether we are dealing with a ***Completely Randomised Design*** or a ***Randomised Block Design***.

EXERCISE 2.2.1

Please refer to standard text books and answer the following questions:

1. Write short notes on the following:
   - i) Distinguish between a Completely Randomized Design and a Randomised Block Design
   - ii) Which of the above design is suitable for splitting two sources of variation in the dependent variable?
   - iii) State the functional form of the models in each

2. Define the following terms:
   - i) Experiment
   - ii) Experimental Error
   - iii) Randomization
   - iv) Replication
   - v) Response variables
   - vi) Extraneous variables
   - vii) Treatment variables

3.
   - i) State the difference between a fixed effect model and a random effects model
   - ii) Describe the importance of the Duncan Multiple range Test to a scientist
   - iii) State the assumptions of ANOVA

Session 2: One Way ANOVA (ANOVA for the Completely Randomised Design)

Introduction: In the Completely Randomised Design, each study unit in each of the multi-categorical comparison groups is an individual study unit. There are no stratifications in the different categories. In this design, treatments are applied to an individual study unit, and measurements are conducted on each. In this type of study design, we conduct what is known as "One-way Analysis of Variance". This method will be discussed in detail in a stepwise process in the steps that follow:

Session Topics: The following topics will be covered:

a. Description of Data and Assumptions
b. The Model
c. The ANOVA Hypothesis
d. Computation formulae for the Completely Randomised Design (CRD)
e. Summarising the Calculations and interpreting findings
f. Application of One Way ANOVA
g. Multiple Comparisons
h. Advantages and Disadvantages of the CRD

Session Objectives:

By the end of this Session, the MPHO should be able to:

1. Examine data sets to discriminate those akin to the Completely Randomised Design
2. State the assumptions of One-way ANOVA
3. State and describe the model for one-way ANOVA
4. Identify the different sources of variation in One-way ANOVA and use them to derive the formulae for one-way ANOVA
5. Summarise One way ANOVA calculations in the ANOVA Table and calculate the test statistic
6. Use the F statistic to draw conclusions of the relationships between the variables in ANOVA
7. Conduct Multiple comparisons to detect least significant paired differences in means in ANOVA
8. Evaluate the CRD to list the advantages and disadvantages of One-way ANOVA

a. Description of Data and Assumptions

	Treatment					
	1	**2**	**3**	……..	**K**	
1.	X_{11}	X_{12}	X_{13}		X_{1K}	
2.	X_{21}	X_{22}	X_{23}		X_{2K}	
3.	X_{31}	X_{32}	X_{33}		X_{3K}	
4.	X_{41}	X_{42}	X_{43}		X_{4K}	
	O	O	O		O	
	O	O	O		O	
	O	O	O		O	
n.	$X_{n1}1$	$X_{n2}2$	$X_{n3}3$		$X_{nk}K$	
Total	$T_{.1}$	$T_{.2}$	$T_{.3}$		$T_{.K}$	$T_{..}$
Mean	$\overline{X_{.1}}$	$\overline{X_{.2}}$	$\overline{X_{.3}}$		$\overline{X_{.K}}$	$\overline{X_{..}}$

Notation:
Let:

- X_{ij} : denotes the values of the measurement for the i-th unit, subjected to the j-th treatment
- T_{*j} : denotes the total of all the units subjected to the J-th (Column Total) treatment
- T^{**} : denotes the Grand Total

Where:

- T_{*j} (the Column Total) is given by:-
  - $T.j = \sum_{i=1}^{r} Xij$
- T^{**} (the Grand Total) is given by:-
  - $T.. = \sum_{j=1}^{C} T.j$; This function is also given by:
  - $T.. = \sum_{j=1}^{C} \sum_{i=1}^{r} Xij$
- The overall Mean $\bar{x}..$ (μ) is given by:-
  - $\frac{T..}{R_c} = \frac{T..}{N}$

The Assumptions of One-Way ANOVA: In order for us to conduct a valid ANOVA, certain assumptions should hold true. They are:-

(a) The K sets of observed data constitute K independent random samples from the respective populations
(b) Each population from which the samples are drawn is normally distributed, with mean μ_j and variance σ^2_j
(c) Each population has the same variance i.e. $\sigma^2_1 = \sigma^2_2 = \ldots\ldots = \sigma^2_k$
(d) The treatment effects (α_j) are known constants and $\sum(\alpha_j) = 0$

b. The Model
The model that is used to predict variation in AVOVA is given by:

$$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

Where:

X_{ij}	=	The value of any given measurement
μ	=	The Overall mean
α_{ij}	=	The effect due to the i-th treatment
ε_{ij}	=	The error component

Interpretation of the Model: The above model simply means that any given measurement in a particular sample is given by: The overall mean, plus the effect due to the treatment, plus a random error component. Any given measurement inherently contains the mean, to which is added the

treatment effect and then an error term due to other effects that are not prescribed in this study. It follows then that we can predict one term if we know the rest.

c. The ANOVA Hypothesis

The Hypotheses can be stated in two ways:

H_O: $\alpha_1 = \alpha_2 = \ldots\ldots\ldots = \alpha_k$
H_A: $\alpha_1 = \alpha_2 = \ldots\ldots\ldots = \alpha_k$

All treatments have a similar effect

Or:

H_O: $\alpha_j = 0; j = 1, 2, \ldots\ldots, c$
H_A: $\alpha_j = 0; j = 1, 2, \ldots\ldots, c$

All treatments do not cause any effect

By looking at the model we observe that our interest is restricted to the k treatments represented in our experiment. Any inferences that we make therefore apply only to these treatment. We do not with to extend our inference to the universal set of possible treatments. When we place such a restriction on our inference goals, we then refer to our model as the "fixed effects model".

Assumptions of the fixed effects model: In order for the parametric procedure of ANOVA to hold as a robust statistical test, there are certain conditions that must be met about the data. If not, then we either have to transform the data, or perform non-parametric tests as will be described later.

d. Computation formulae for the Completely Randomised Design (CRD)

Consider the data-set given below:

	Independent Variable (Treatment or Intervention)			
Dependent Variable	**Participatory**	**Lecture**	**Control**	**Overall**
Study units				
Student 1	49.7	42.2	2.3	
Student 2	48.9	35.8	4.8	
Student 3	50.5	46.7	7.4	
Student 4	45.4	39.7	4.7	
Student 5	52.8	35.5	4.7	
Student 6	42.1	21.1	5.1	
	289.4	227.3	29.0	545.7
Mean	**48.2**	**37.0**	**4.9**	**26.0**

From the above dataset, we can observe 3 types of variation:

1. Between Groups variation: This is the first important source of variation that we can calculate. It is obtained by taking the group means as variables, and calculating the sum of their squared deviations **from the overall mean**. E.g. The Participatory Group Mean (48.2) minus the Overall Mean (26.0). We

then square all the deviations of the group means from the overall mean and sum then up: $[\sum (\bar{x}_j - \mu_X)^2]$.

This calculation is also a measure of variance and is called the **Among-Groups Variation** or the Column sum of Squares, or **SSC**. It is expressed by the following notation: -

$$SSC = n\sum_{i=1}^{C}\left(\overline{X}.j - \overline{X}_{..}\right)^2$$

$$SSC = \sum_{j=1}^{C}\frac{T^2.j}{r} - \frac{T^{2..}}{rc}$$

$$SSC = \sum_{j=1}^{C}\frac{T^2.j}{r} - \frac{T^{2..}}{N}$$

2. The Total Sum of Squares: This is the second important source of variation. We can calculate the sum of all squared deviations for **each** individual observation, this time from the **Overall mean** (μ_X) of the entire data set and not the group mean like we did with the Error Sum of Squares. It can be expresses as: $[\sum (X_{ij} - \mu_X]$. For instance, if a variable like the Senior One score in the participatory arm is considered, from it we subtract the overall mean (26.0) and square the result $(49.7 - 26.0)^2$. When we sum up all the squared deviations from the overall mean is calculated, we obtain what we call the **Total Sum of Squares**. It is expressed by the following notation:-

$$\mathbf{SST} = [\sum (\bar{x}_j - \mu_X)^2]$$

$$SST = \sum_{j=1}^{C}\sum_{i=1}^{r}\left(X_{ij} - \overline{X}_{..}\right)^2$$

$$SST = \sum_{j=1}^{C}\sum_{i=1}^{r}X^2ij - \frac{T^2..}{rc}$$

$$SST = \sum_{j=1}^{C}\sum_{i=1}^{r}X^2ij - \frac{T^2..}{N}$$

3. Within Groups Variation: The variation of each individual measurement from its group mean, e.g. the senior 1 score for the participatory arm (49.7), if subtracted from the mean Participatory arm performance (48.2) gives a measure of -1.5. Similarly, the Senior 4 score (45.4) for the participatory arm (45.4), when subtracted from the participatory arm mean gives a measure of 2.8. We can calculate this for every measurement in a given group ($X_{ij} - \bar{x}_j$). However, if we simply sum up all these calculations, we obtain a sum of zero, because they are deviations from the mean. We can square all the calculations, so that we eliminate the plus or minus sign $[(X_{ij} - \bar{x}_j)^2]$ and then sum up all the squared calculations $[\sum (X_{ij} - \bar{x}_j)^2]$. What we then obtain is the sum of squared deviations from the group mean. It is indeed a measurement of variance for the group and is referred to as the "Within Group Variation".

We then sum up all these calculations for all the various "treatment" groups, in this case – Participatory, Lecture and Control. What we obtain is called the **Error Sum of Squares** or **Sum of Squares due to Error**, or **SSE**.

$$SST = \sum_{j=1}^{C}\sum_{i=1}^{r}\left(X_{ij} - \overline{X}.j\right)^2$$

NB: As you can see, this is the most cumbersome parameter to calculate of all the three sources of variation. We have to painfully get through the process of calculating the squared deviations of each value from its group mean and then sum up all these. There is however a shortcut to this process. Therefore:

Since: $SST = SSC - SSE$

Then: $SSE = SST - SSC$

e. Summarising the Calculations and interpreting findings

Summarising the computations: After the calculations, we can summarise the findings in what is called the ANOVA Table:

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**F-Statistic (Computed) (Fc)**
Between Groups (Column means)	Column Sum of Squares (SSC) $SSC = \sum_{j=1}^{C}\frac{T^2.j}{nj} - \frac{T^{2..}}{N}$	$(C-1)$ (Total number of columns minus one)	$S^2_1 = \frac{SSC}{C\text{-}1}$	$\frac{S^2_1}{S^2_2}$
Within Groups (Error)	Error Sum of Squares $SSE = SST - SSC$	$C(R-1)$ Number of Columns X Rows minus 1*	$S^2_2 = \frac{SSE}{C(R\text{-}1)}$	
Total	Total Sum of Squares $SST = \sum_{j=1}^{C}\sum_{i=1}^{r} X^2 ij - \frac{T^2..}{N}$	$(RC-1)$		

- Note here that the function C(R-1) is indeed the total number of observations minus one. This is in cases where there are equal numbers of respondents in each category. However, there are situations in which the treatment groups do not necessarily have the same numbers. In this case, we use N-1.

Decision rule: Reject H_0 if the computed value of F ($F_{computed}$) is greater than the Tabulated value F ($F_{tabulated}$) at Degrees of Freedom C-1; C(r-1). Notice that F will be large if MSB is much larger than MSE. Thus if F is large, this means that the variation in SST attributable to ***differences between the groups*** is much larger than that attributable to ***differences within the groups***; whereas if F is small, it means that the variation in SST attributable to ***differences between the groups*** is much smaller than that attributable to ***differences within the groups***.

NB: The above findings are true for One Way ANOVA.

f. Application of One Way ANOVA

EXAMPLE 1: The Data in the table below represents the number of hours on pain relief provided by five different brands of head-ache tablets administered to 25 patients. The 25 subjects were randomly divided into 5 groups and each group was treated with a different brand.

	Tablet					
	A	**B**	**C**	**D**	**E**	
	5	9	3	2	7	
	4	7	5	3	6	
	8	8	2	4	9	
	6	6	3	1	4	
	3	9	7	4	7	
Total	**26**	**39**	**20**	**14**	**33**	**132**
Mean	5.2	7.8	4.0	2.8	6.6	5.28

Perform ANOVA and test the hypotheses at the 0.05 level of significance that the number of hours of relief provided by the tablets is the same for all five brands.

Solution: We follow the following steps:-

1. Display and describe the data: As displayed above
2. Assumptions: The assumptions of One way ANOVA hold

   The model: $\mathbf{X_{ij} = \mu + \alpha_j + \varepsilon_{ij}}$

   Where:

   μ = The Overall mean

   α_j = The effect due to the i-th treatment

   ε_{ij} = The error component

3. Hypotheses: H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

   H_A: At least two of the means are not equal

4. Distribution of the test statistic: The F-Distribution
5. Decision Rule: The critical region: Reject at the 0.05 level of significance if FC>FT at degrees of freedom (5-1) = 4; (25-5) =20
6. Computation:

$$SST = (5^2+4^2 \ldots\ldots\ldots+7^2) - \frac{132^2}{25}$$

$$= 834 - 696.96$$

$$=137.04$$

$$SSC = (26^2+ 39^2+ \ldots\ldots\ldots+ 33^2) - \frac{132^2}{25}$$

$$= 776.4 - 696.96$$

$$= 79.44$$

$$SSE = 137.04 - 79.44$$

$$= 57.6$$

The results of the computation are exhibited in the table below:

ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic (Computed) (Fc)
Column means	79.44	4	79.44/4 = 19.860	$\frac{19.860}{2.880} = 6.90$
Error	57.60	20	57.60/20 = 2.880	
Total	137.04	24		

Decision: At α = 0.05, and degrees of freedom 4, 20, the Tabulated F-Value is 2.87. Since the Computed F value Fc is greater than the tabulated value, we reject H_0

Conclusion: We conclude that the mean number of hours of relief provided by head-ache tablets is not the same for all five brands.

EXAMPLE 2:

Suppose that there are 4 different forms of contraceptives. A family health worker has received complaints from a number of his clients that they have gained weight as a result of taking the contraceptives. The family health worker has also noticed this, and is aware that the gain in wait varies from individual to individual. He does not however know if the variation is due to genetic differences in his clients or due the different types of contraceptives being used. We therefore want to test the following hypothesis:

H_o: There is no difference in the average weight gain between groups using the different contraceptives i.e. $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_A: The average weight gain of at least one of the μ_i's is different from the another i.e. $\mu_i \neq \mu_j$ where $i \neq j$

Let y_{ij} represent the weight gained by the j^{th} client using the i^{th} contraceptive,

Where j = 1, 2, ... n (n = # of clients using the i^{th} contraceptive)
Where i = 1, 2, ... m (g = # of different contraceptives - groups)

	Type of Contraceptive used				
	1	**2**	**3**	**4**	
Weight gained	12.2 9.5 11.6 13.0 10.1 9.6	4.9 10.6 7.0 8.3 5.5 11.7	8.0 12.1 5.7 8.6 7.2 12.4	4.6 6.7 5.0 3.8 8.2 7.7	n = **6** g = **4** $y_{\bullet\bullet} = \sum_{i=1}^{4}\sum_{j=1}^{6} y_{ij}$ = **204** $\sum_{i=1}^{4}\sum_{j=1}^{6} y_{ij}^2$ = **1912.7** $\sum_{i=1}^{4} y_{i\bullet}^2$ = **10,872**
Sum ($\sum_{j=1}^{6} y_{i\bullet}$)	66.0	48.0	54.0	36.0	
Average ($\bar{y}_{i\bullet}$)	11.0	8.0	9.0	6.0	

The total sum of squares in this data is given by:

$$\sum \sum (y_{ij} - \bar{y}_{\bullet\bullet})^2$$

Where $\bar{y}_{\bullet\bullet}$ is the overall mean weight gain of all the clients? The above is the Sum of Squares Total (SST). We want to divide this total variation into:

a) Variation Within each group of the same type of contraceptive users. This is known as the Sum of Squares within (SSW) Groups.

b) Variation Between the different groups of contraceptive users, i.e. Sum of Squares between (SSB) groups;

So that SST = SSB + SSW

With a little mathematical manipulation, it can be shown that:

$$\sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2 = \sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2 + n\sum_{i=1}^{g}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$$

SST SSW SSB

Computational Form:

$$\text{SST} = \sum_{i=1}^{g}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2$$

$$= \sum_{i=1}^{g}\sum_{j=1}^{n} y_{ij}^2 - (\bar{y}_{\bullet\bullet})^2 / N = 1912.7 - (204)^2/24 = 178.7$$

SSB = $n\sum_{i=1}^{g}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$

= $\sum_{i=1}^{g} y_{i\bullet}^2 \Big/ n - (y_{\bullet\bullet})^2 / N$ = 10,872/6 - $(204)^2$/24 = 78.0

SSW = $\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2$

= SST - SSB = 178.7 - 78.0

= 100.7

For the above example therefore, the ANOVA table would be

ANOVA TABLE

Source of Variation	df	Sum of Squares (SS)	Mean Square	F-Ratio df = m-1, m(n-1)
1. Within Groups (error or residual)	20	SSW = 100.7	MSW = 100.7/20 = 5.035	F_c = 26/5.035 = 5.16 ($F_{3, 20, 0.05}$ = 3.10)
2. Between group	3	SSB = 78.0	MSB = 78/3 = 26	($F_{3, 20, 0.01}$ = 4.94)
Total	23	TSS = 178.7		

Interpretation of ANOVA Results

Since the calculated F_c is greater than the tabulated $F_{(g-1),(n-1),(\alpha)}$, the mean squares between groups are significantly bigger than the mean squares within groups. We therefore reject the null hypothesis and conclude that the mean weight gain in at least one of the groups is significantly different from at least one other group.

Of course this does not tell us which groups are significantly different from each other. To find out which ones are different from each other, we apply the t-test or other multiple comparison tests. But obviously the group with the biggest mean (Group 1) must be significantly different from the group with the smallest mean (Group 4).

g. Multiple Comparisons

ANOVA is a powerful procedure for testing homogeneity of a set of means. However, it has one shortfall: if the Null Hypothesis is rejected, and we conclude that the means are in fact not equal, we still do not know which means in the different categories differ significantly from each other. We do not know which ones of the different category means are not equal. There are several statistical procedures that can assist us to conduct multiple comparisons or Post Hocs as they are known. They include:

- The Duncan Multiple Range test (Least Significant range method)
- The Turkey's HSD Test (Highest Significant Test)

In this Session, we shall concentrate on only one method – The Multiple range test of Duncan.

Procedure: Let us assume that the ANOVA procedure has led to a rejection of the null-hypotheses of equal population means. It is also assumed that the **k**-range samples are all of the same size, **n**. The range of any subset of **p** sample means must exceed a certain value before we consider

any of the p population means to be different. This value is called the Least Significant Difference or Range (LSD) for the **p**-means, and is denoted by **R p**.

Where: $Rp = r_p * S_{\overline{X}} = r_p\sqrt{S^2/n}$

The Sample Variance **S²**, is the estimate of the common variance **σ²,** and is obtained from the **Error Mean Square** in the ANOVA Table.

The term r_p is called the ***Least Significant Studentised Range***. It depends on the ***desired level of significance***, and the ***degrees of freedom*** of the **Error Mean Square**.

The values can be obtained from r-tables, for p=2, 3, 4 ….10 means.

Example: Reconsider the example given above, in which different mean hours for relief of headaches were compared for different treatments. Remember that we rejected the null hypothesis, and therefore, we accepted the alternative hypothesis that the tablets did not have the same average hours of relief. We are now interested in finding out which ones of these means are different. We have to do paired post-hoc comparisons, and the test we have chosen is the Multiple Range Test by Duncan. We conduct the following procedures:

1. We arrange the sample means in increasing order of magnitude

$\overline{X_4}$	$\overline{X_3}$	$\overline{X_1}$	$\overline{X_5}$	$\overline{X_2}$
2.8	4.0	5.2	6.6	7.8

2. We obtain ***S²***. Remember, it is the Error Mean Square; from the ANOVA table, it is 2.88
3. We Obtain the values of r_p at degrees of freedom of the Error mean Square (20) and α=0.05. We have to determine the 4 values, for P= 2, 3, 4 and 5. We multiply each of these r_p with the term $\sqrt{S^2/n}$ to obtain the corresponding Rp, which we use for the multiple comparisons. In this case, the term $\sqrt{S^2/n} = 0.76$.

P	2	3	4	5
r_p	2.95	3.097	3.190	3.255
R_P	2.24	2.35	2.42	2.47

4. We then conduct the multiple comparisons and make statistical deduction for each.

    1. Since $\overline{X}_2 - \overline{X}_5$ = 7.8 – 6.6 = 1.2, and this is less than the R_{P2} (2.24), then we conclude that $\overline{X_2}$ and $\overline{X_5}$ are **not significantly different**
    2. Since $\overline{X}_2 - \overline{X}_1$ = 7.8 – 5.2 = 2.6, and this is greater than the R_{P3} (2.35), then we conclude that $\overline{X_2}$ and $\overline{X_1}$ **are significantly different**
    3. Similarly $\overline{X}_2 - \overline{X}_3$ = 7.8 – 4.0 = 3.8, and this is greater than the R_{P4} (2.42), then we conclude that $\overline{X_2}$ and $\overline{X_3}$ **are significantly different**

4. Similarly $\overline{X}_2 - \overline{X}_4$ = 7.8 – 4.0 = 3.8, and this is greater than the R_{P5} (2.47), then we conclude that $\overline{X_2}$ and $\overline{X_4}$ **are significantly different**
5. $\overline{X}_5 - \overline{X}_1$ = 6.6 – 5.1 = 1.4, and this is less than the R_{P2} (2.24), then we conclude that $\overline{X_1}$ and $\overline{X_5}$ are **not significantly different**
6. $\overline{X}_5 - \overline{X}_3$ = 6.6 – 4.0 = 2.6, and this is greater t than the R_{P3} (2.35), then we conclude that $\overline{X_3}$ and $\overline{X_5}$ are **significantly different**
7. $\overline{X}_1 - \overline{X}_3$ = 5.2 – 4.0 = 1.2, and this is less than the R_{P2} (2.24), then we conclude that $\overline{X_1}$ and $\overline{X_3}$ are **not significantly different**
8. $\overline{X}_1 - \overline{X}_4$ = 5.2 – 2.8 = 1.4, and this is less than the R_{P2} (2.35), then we conclude that $\overline{X_1}$ and $\overline{X_4}$ are **not significantly different**
9. And so on and so forth

We note here that the manual computation is cumbersome, potentially confusing and tedious. Fortunately, computers can do all the multiple comparisons simultaneously with the ANOVA if commanded and the demonstration above is only for you to understand the rationale. The end result of a post-hoc test is to determine which group means are significantly different from others.

h. Advantages and Disadvantages of the CRD

Advantages of the Completely Randomised Design (One-Way ANOVA)

- Simplicity in the lay-out of the design
- Statistical analysis and interpretation of the results are straight forward
- It does not necessarily require the use of equal sample sizes for each treatment
- It allows for maximum number of degrees of freedom for the errors
- It does not require the experimental units/subjects to participate under more than one treatment level

Disadvantages:

- It does not cater for the effect due to the different treatments; we need to conduct further tests to decide which groups differ significantly
- Its assumptions mean that we have to conduct tests to ensure that ANOVA is relevant
- Variances of the different treatment groups should be homogenous

2.7 Unit 7: LINEAR REGRESSION AND CORRELATION

2.7.1 Introduction to the Unit

In several statistical analyses we looked at methods to describe the association or relationships between two variables. In this unit, we shall go beyond this: we shall look at *methods of using available information to predict the value of one characteristic using another*: i.e. by knowing X, we can predict Y.

Illustration: We know that children under two years are rapidly growing. We can therefore predict the "average normal" weight of a child at a given age: If we know the age (X) of a child in months, we can predict their likely weight (Y). To do this, we study several sets of data from normal populations, and come up with a mathematical model that helps us predict with a certain level of accuracy the likely values of weight, for a given age, and the average of these values. These approaches are widely used in the health sciences and they are generally called "Statistical Modelling".

One of the commonly used methods of statistical modelling is linear regression and correlation analysis. These methods will be the subject of this unit. After determining whether there is a relationship between the two variables, we can test two other issues:

1. Is the relationship we are assuming (e.g. a linear relationship) significant in its prediction of one variable from another?
2. If significant, how strong is the relationship?

When we use a linear model (meaning that we assume that age is linearly related to weight in the above illustration), then we refer to Linear regression and correlation.

Types of variables: We need two types of variables:

(1) The Independent or Explanatory Variable
(2) The Dependent or Response variable or Outcome variable

This means in simple terms that one variable is used as the basis for predicting the other. It is usually the less costly or less difficult one to measure, and the statistical methods of linear regression or Correlation are then applied to predict the other variable, without necessarily going through the cumbersome process of measuring it. Correlation is mainly about predicting a relationship, while Regression is about predicting a value from another.

Illustration: Please note the following scenarios: -

- In a certain field trial of communication approaches for Reproductive Health Education in Secondary Schools, the researchers measured baseline knowledge of Reproductive health. It was noted that the average performance increased with class. We can say therefore that Baseline knowledge of RH Correlates with Class level of a student. We note that both class and baseline knowledge score are continuous variables. However, class is "Fixed" (A Discrete Random Variable), yet Knowledge Score is continuous (A Continuous Random Variable).

- In the example given on age and weight of a child, we are interested in predicting the normal weight of a child for a given age. We therefore use the method of correlation, because both "age" and "weight" are continuous random variables. If we fix the independent variable "age", then we can do linear regression analysis.

2.7.2 Unit Outline

The following Sessions will be covered:

Session 1: Linear Regression
Session 2: Correlation Analysis
Session 3: Multiple Linear Regression
Session 4: Collinearity and Multi-collinearity

2.7.3 Instructional goal

The MPHO should be able to apply and appraise the methods of linear regression and correlation in predicting outcomes related to stated fixed variables

2.7.4 Unit Objectives

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

1. Select appropriate variables and use Linear Regression methods to predict one variable from another
2. Employ the method of correlation to investigate the relationship between two continuous variables
3. Construct predictor models for several variables using multiple linear regression and evaluate the models for possible confounding
4. Describe the phenomenon of collinearity and multi-collinearity

2.7.5 Time Frame

1 WEEK

2.7.6 Content

Session 1: Simple Linear Regression

Introduction: It involves determining whether there is indeed a linear relationship between one or more independent variables (X_1, X_2, X_3 … etc) and a dependent variable (Y) e.g., the variation of Age, Height and Blood Pressure (Independent Variables) with Weight (the Dependent Variable).

We do this by determining a linear equation for predicting the value of one outcome variable for a given independent (explanatory) variable. We then test the equation to assess if the linear relationship is statistically significant, and its strength in assisting us to predict values. The variables are expressed as earlier stated:

- **X** is the Independent variable (Explanatory Variable)
- **Y** is the Dependent variable (Outcome Variable)

Session Topics: The following topics will be covered:

a. Definition and meaning
b. Simple Linear Regression
c. Summary of the Steps in Regression Analysis
d. Step 1: Assumption of probable Linear Relationship
e. Step 2: The Equation for the Line of Best Fit
f. Step 3: Evaluating the strength of the relationship
g. Step 4: Using the Regression Equation

Session Objectives:

By the end of this Session, the MPHO should be able to:

1. Define and explain the meaning of Linear Regression and Simple Linear Regression
2. Employ appropriate plots and diagrams to demonstrate the Assumption of Linearity between two variables
3. Solve for the line of best fit between two variables that exhibit a linear relationship, with the help of the Least Squares Criteria
4. Evaluate the strength of linear relationships using the Coefficient of Determination, ANOVA or the t – test
5. Appraise the role of linear regression equations in Prediction and Estimation

a. Definition and meaning

Regression is a statistical tool for evaluating the relationship of one or more independent variables X_1 ,X_2, X_3..... X_{k-1}, .X_K to a single continuous dependent variable. Examples of continuous variables include; height, age, blood pressure heart rate and medical care expenditure.

You will find out from your readings that one always has to be cautious about the results obtained from regression analysis or more generally from any form of analysis seeking to quantify any association between two or more variables. Although the statistical computations used to produce an estimated measure of association maybe correct, the estimate itself may be biased. Such a bias may result from the methods used to select subjects for the study, errors in the information used in the statistical analysis or even variables other variables that cannot account for the observed association.
It is important to note that the finding of a statistically significant association in a particular study does not establish a causal relationship.

To evaluate claims of causality, one must consider criteria that are external to specific characteristics and results of any single study. Let us now review a list of general criteria for assessing the extent to

which available evidence supports a causal relationship as formalized by Bradford Hill in 1971. The list contains seven criteria, which we summarize below:

- Strength of the association
- Dose-response
- Lack of temporal ambiguity
- Consistency of findings
- Biological and theoretical plausibility of the hypothesis
- Coherence of evidence
- Specificity of the association

For more details about those seven criteria, refer to Applied Regression Analysis and other Multivariate Methods by Kleinbaum et al, Second edition page 39.

b. Simple Linear Regression

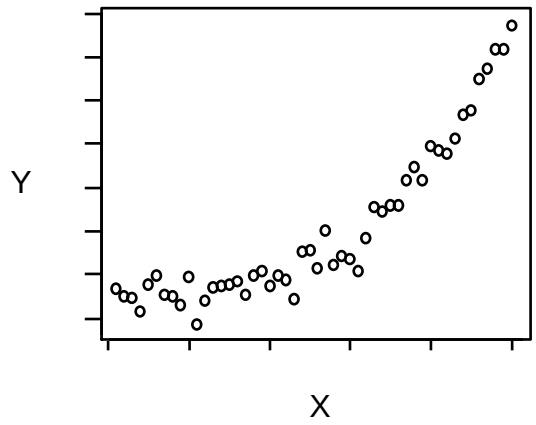
This is the simplest form of the general regression problem which deals with **one** dependent variable (Y) and **one** independent variable (X). It is denoted by the equation:

$$Y = \beta_0 + \beta_1 X$$

Where: β_0 is the intercept and β_1 is the slope.

Let's begin this section by describing a step-by-step strategy used in fitting a regression model

1- Begin by assuming that a straight line is the appropriate model for your data. (Linearity can be investigated). The figure below shows a scatter diagram of X and Y



2- Find the best straight line that best agrees with your data
3- Determine whether the straight line found in step 2 significantly helps to describe the dependent variable Y (check certain basic assumption e.g., normality).
4- Examine whether the assumption of straight line model is correct
5- If assumption of straight line is found to be invalid in step 4, fit a new model
6- Continue to try new models until one is found

Assumptions: For us to conduct a regression of Y on X, the following assumptions are pertinent:

1. Values of the independent variable X are said to be "fixed" i.e. non random. For instance, the "Age" for which we are determining the possible weight must be fixed e.g. "at 2months" or "3 months" and not a continuum of ages.
2. The Variable X is measured with a negligible magnitude of error.
3. For each value of X, there is a sub-population of Y values (e.g. for each age X, there is a distribution of values of Weight: - if we take a random sample of 100 normal six month old children, not all of them will have the same weight but most of them will have a weight range of say 6 to 8Kg). These subpopulations of Y values should be normally distributed – This is called the ***Assumption of Normality***.
4. The variances of the sub-populations of Y are equal:
   - If X_1 has a subpopulation of Y_A,
   - And X_2 has a subpopulation of Y_B,
   - Then the variances of each of these sub-populations should be equal.

   This is called the ***Assumption of Variance***.
5. The means of the subpopulations of Y all lie on the same straight line. This is called the ***Assumption of Linearity***. We can write an equation that represents this line as follows:

   $$\mu_{Y/X} = \beta_0 + \beta_1 X$$

   Where:

   $\mu_{Y/X}$ = The mean of the Sub-population of Y -Values for a given value

   β_0 and β_1 = Population regression Coefficients

   X = A given fixed value of the independent variable

   Geometrically:

   β_0 = Y - Intercept

   β_1 = The Slope

6. The Y values are statistically independent i.e. in drawing a sample the values of Y chosen at a particular value of Y are in no way dependent on the values of Y chosen at another value of Y. This is the ***Assumption of Independence***

The Regression Model: The assumptions are summarized by means of the following equation which we call the regression model:

$$y = \beta_0 + \beta_1 + \varepsilon$$

Where y is the typical value from one of the subpopulations of Y, β_0 and β_1 have already been defined and ε is the error term. From the above equations we observe that,

$$\varepsilon = y - (\beta_0 + \beta_1 X)$$

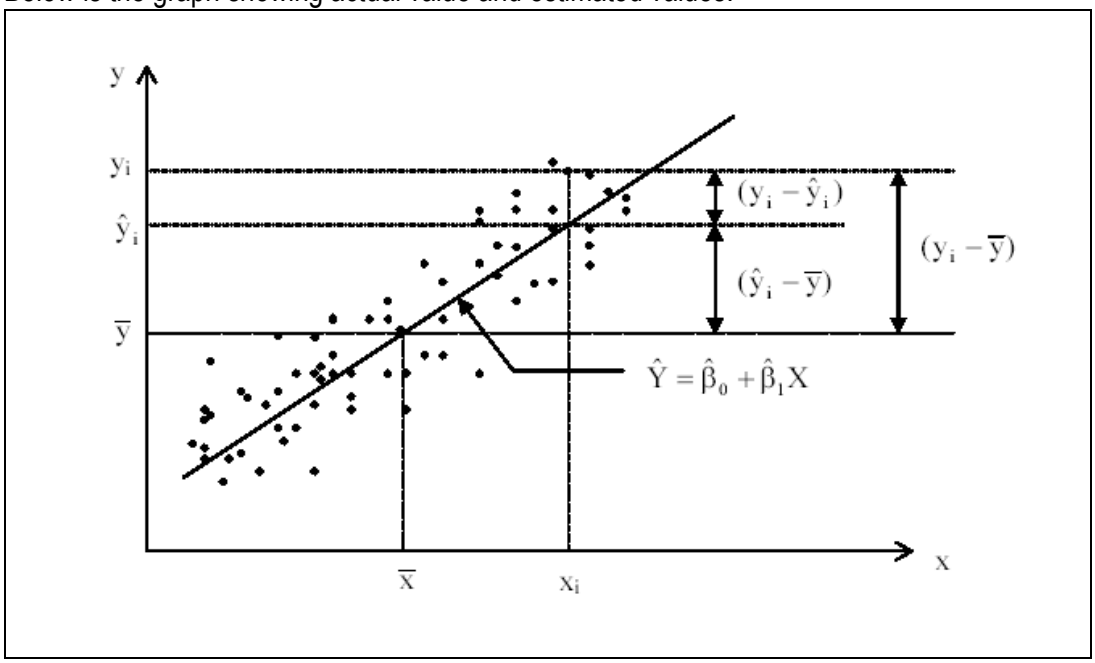
$$\varepsilon = y - \mu_{Y/X}$$

$\mu_{Y/X}$ is defined also as the expected value of Y given, or the mean of the subpopulation of values of Y for a given value of X. Errors that are obtained by subtracting estimated values from the actual values are important. There are assumptions that are assumed to be associated with them:

- Errors are independent of each other

- Errors have zero mean and a common variance
- Errors follow a normal distribution.

Below is the graph showing actual value and estimated values.

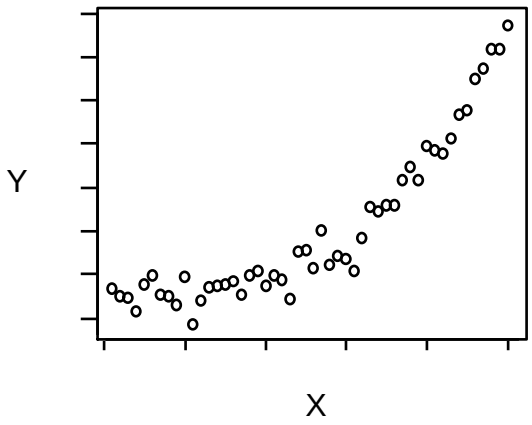


c. Steps in Regression Analysis

1- Determine whether there is a probable linear relationship and whether the assumptions of linearity are met in the data available for analysis
2- Obtain the equation of the line that best fits the sample data
3- Evaluate the equation to get an idea of the strength of the relationship
4- If the data conform satisfactorily to a linear model, use the equation to **Predict** and to **Estimate**
    a. **Predict** – the value of likely Y for a given value of X
    b. **Estimate** – the mean of the subpopulation of Y values assumed to exist at a given value of X

d. Step 1: Assumption of probable Linear Relationship

We use various data presentation methods to demonstrate the assumption of a probable linear relationship. One of these is a scatter-plot. From the scatter-plot, we can then infer that there in a linear trend. However, we cannot directly predict the equation of best fit, because any two people drawing a line will draw it differently.
e.g.:



e. Step 2: The Equation for the Line of Best Fit

Methods of estimating the parameters: In order to obtain both the slope and intercept, number techniques which minimize errors are used. They include; Least-Squares, minimum variance and maximum likelihood methods.

Activity: *Read Applied Regression Analysis and other Multivariate Methods by Kleinbaum et al, Second Edition pages 49-53 for further details*.

The Least Squares Method: Here, we shall describe one of the most commonly used methods – the ***Least Squares Criterion***. The resulting line therefore is the ***Least Squares Line***. Since the Equation is generally written as:-

$$Y = \beta_0 + \beta_1 X$$

There are two important constants that we need to predict. They are:-

β_0: The Y - Intercept

β_1: The amount by which Y changes for each unit change in X . It is also called the slope of the line.

If we have these two constants, we can substitute various values of X in the equation, to obtain the corresponding **given** value Y . Any two coordinates will also enable us to draw the line of best fit. β_0 and β_1 can be obtained by the solution of two simultaneous equations below:

$$\sum Yi = n\beta_0 + n\beta_1 \sum Xi \quad \text{..... (i))}$$

$$\sum XiYi = \beta_0 \sum Xi + \beta_1 \sum Xi^2 \quad \text{..... (ii)}$$

Example: The following anthropometric data was collected for 10 children under one year:

Observation	Age	Weight
1	10	11
2	7	10
3	10	12
4	5	6
5	8	10
6	8	7
7	6	9
8	7	10
9	9	11
10	10	10

Find the equation of the line of best fit that describes the relationship of Age to Weight.

Solution: We construct a table that summarises the data and the important parameters. This table also enables us to display and describe the dataset as well as calculate the other parameters that will be useful in the equations that will enable us determine the important parameters for the Least Squares Line. These are summarised in the table below:

Observation	**X**	**Y**	**X^2**	**Y^2**	**XY**
1	10	11	100	121	110
2	7	10	49	100	70
3	10	12	100	144	120
4	5	6	25	36	30
5	8	10	64	100	80
6	8	7	64	49	56
7	6	9	36	81	54
8	7	10	49	100	70
9	9	11	81	121	99
10	10	10	100	100	100
N=10	$\sum Xi$ =80	$\sum Yi = 96$	$\sum Xi^2 = 668$	$\sum Yi^2 = 952$	$\sum XY = 789$

$$\sum Yi = n\beta_0 + n\beta_1 \sum Xi \quad \text{..... (i)}$$

$$\sum XiYi = \beta_0 \sum Xi + \beta_1 \sum Xi^2 \quad \text{..... (ii)}$$

Substituting:
(1) $96 = 10\ \beta_0 + \beta_1\ 80$
(2) $789 = 80\ \beta_0 + 668\ \beta_1$

Therefore, when we solve the simultaneous equation:
Let us first use equation 1:

$$96 = 10\ \beta_0 + \beta_1\ 80$$

Therefore: $10\ \beta_0 = 96 - 80\ \beta_1$

Thus: $\beta_0 = \frac{96 - 80\ \beta_1}{10}$

Substituting into the second equation:

$$789 = \frac{80(96 - 80\ \beta_1)}{10} + 668\ \beta_1$$

Therefore: $7890 = 7680 - 6400\,\beta_1 + 6680\,\beta_1$
Collecting Like terms: $7890 = 7680 + 280\,\beta_1$
Thus: $210 = 280\,\beta_1$
Consequently: $\beta_1 = \frac{210}{280} =$ **0.75**

We then substitute and solve for β_0 = **3.6**

$\beta_0 = 3.6$; $\beta_1 = 0.75$

Therefore, the equation for the Least Squares Line is:

$$Y = 3.6 + 0.75X$$

Alternative Formulae for β_0 and β_1: We can also estimate the equation parameters using the least squares criterion. In this method, the slope β_1 is given by:

$\beta_1 = SS_{xy} / SS_x$ With a standard error $se(\beta_1) \approx SS_{xy} / \sqrt{SS_x}$

The sum of squares and cross product are given by:

$$SS_x = \sum (x - \bar{x})^2 = \sum x^2 - (\sum x)^2 / n$$

$$SS_y = \sum (y - \bar{y})^2 = \sum y^2 - (\sum y^2)/n$$

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - (\sum x)(\sum y)/n$$

Therefore: $\beta_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$

The Intercept, β_0 is given by: $\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$ Or: $\beta_0 = \sum \bar{y} - \beta_1 \sum \bar{x}$

With a standard error $se(\beta_0) \approx SS_{xy}\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)$

From the above example, we can substitute into these alternative equations and calculate the slope and intercept as follows:

$\beta_1 = \frac{(10 \text{ X } 789) - (80)(96)}{(10\text{X}668) - (80)^2}$

$\beta_1 = \frac{(7890) - (7680)}{(6680) - (6400)}$

$\beta_1 = \frac{210}{280} =$ **0.75**

Therefore:

$\beta_0 = \frac{96 - (0.75)(80)}{10} = \mathbf{3.6}$

Comments: The Least Squares Criterion results in the best line of fit. It does not pass through all points on the scatter diagram, but most observed points deviate from the line by varying amounts. "The Sum of the squared vertical deviations of each observed data points (Y_i) from the least squares line is smaller than the sum of the squared vertical deviations of the data points from any other line". If we square the vertical distance from each observed point (Y_i) to the Least Squares Line and add these, the resulting total will be smaller than the similarly computed value for any other line that can be drawn through the points: this is the Least Squares Line.

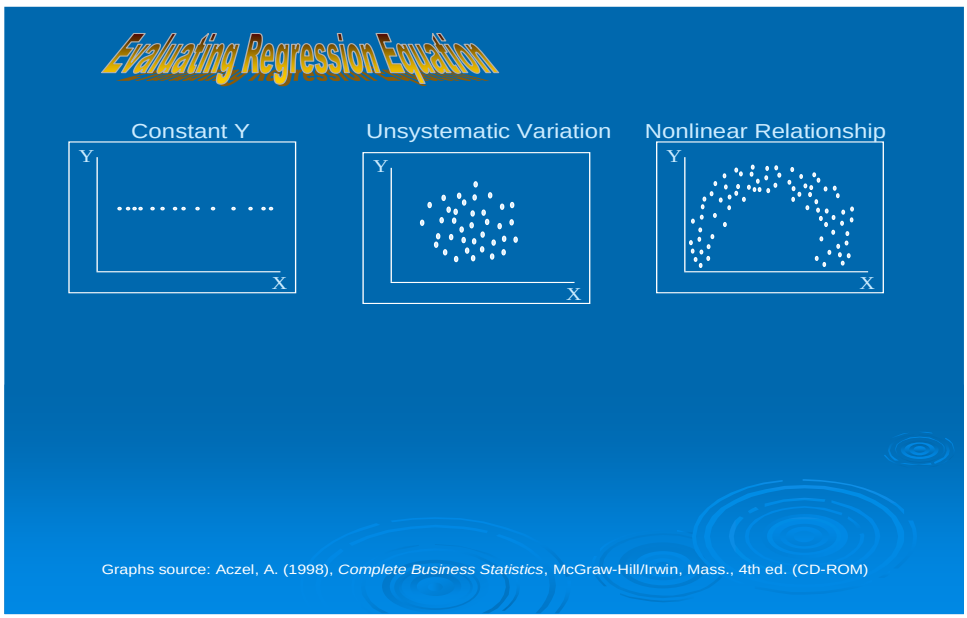
f. Step 3: Evaluating the strength of the relationship

Once the regression equation has been obtained, it must be evaluated to determine whether it adequately describes the relationship between two variables and whether it can be used effectively for prediction and estimation purposes. The first step in evaluating the regression model is to test the following hypothesis:

$H_0: \beta = 0$

$H_a: \beta \neq 0$

If in the population the relationship between X and Y is linear, β the slope of the line that describes this relationship will either be positive, negative or zero. If the slope (β) is zero, data drawn from the population will in the long run yield regression equations that have little or no value for prediction and estimation purposes. The relationship between X and Y could be best described better by a non linear model. This implies that when H_0 is not rejected, even though the relationship between X and Y may be linear, it is not enough for X to be a predictor of Y. Let us look at graphical presentations that can lead to either rejecting or failing to reject a null hypothesis.



When H_0: $\beta_1 = 0$ is not rejected: If the relationship appears linear, yet the slope is zero, it may mean that:-

1. Though the relationship between X and Y is linear, it is not strong enough for X to be of any value in predicting Y
2. The relationship may not actually be linear, but curvilinear, logistic, ovoid etc. We have to explore for another possible model that is not linear

When H_0: $\beta_1 = 0$ is rejected: In this situation, there are two options:-

1. That the regression relationship is linear and of sufficient strength to justify the use of the sample regression equation to estimate Y for a given X , or
2. That there is a good fit to a linear model, but some curvilinear model may provide a better fit.

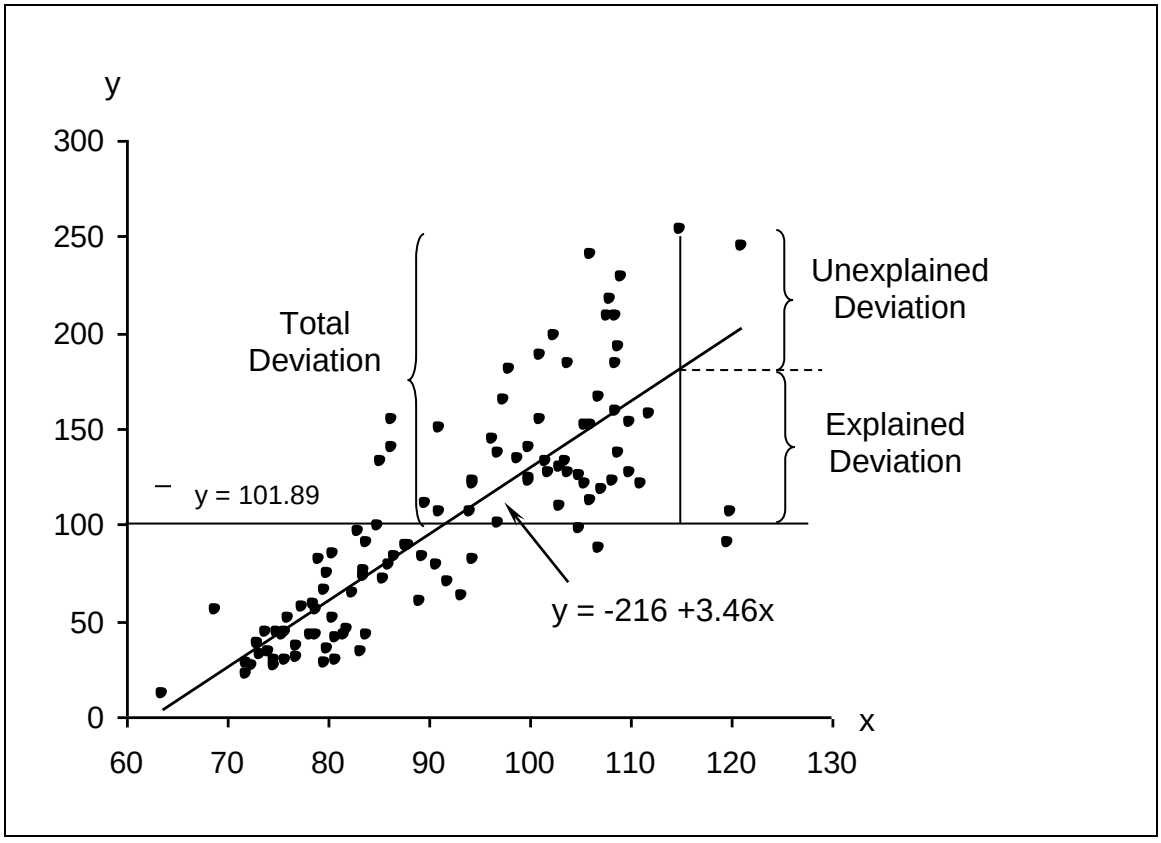
We therefore have to conduct procedures to evaluate the value of the regression equation. We need to:

1. Test Hypotheses about the slope, to gauge whether it is statistically significant. Here, we use:
   a. The t – test , or
   b. The F – Test (ANOVA)

2. Determine the Closeness of fit of the sample regression equation. To do this, we compute the Coefficient of Determination (R^2).
3. To determine the strength of the association: We use the Correlation Coefficient (R)

1. Testing H_0: $\beta_1 = 0$ with the Coefficient of Determination

It is an objective measure of the Strength of the Association – to what extent is the observed variation explained by the regression equation as compared to the un-explained or random variation (or Error). Consider the diagram below:



To derive the Coefficient of Determination, we need to calculate some terms. Consider a given data set that we use to obtain the line of best fit in the figure above. We need to appreciate 3 parameters.

1. Remember for each discrete value of X , there is a sub-population of Y values. The means of the sub-populations of Y values constitute the Linear Regression Line. We can denote each of these means by $\hat{y}$ or the y^{est} (The estimated value of Y using the regression equation.
2. We can also determine the overall mean of all the observations. This may be denoted as $\bar{y}$.
3. Any given value of y_i

We note that the **Total Deviation** of any given value of y_i from the overall mean $(y_i - \bar{y})$ is composed of two parts: The **Explained Deviation** (deviation from the overall mean that is explained by the regression equation) $(\hat{y} - \bar{y})$ and the **Unexplained Deviation** (deviation from the overall mean, that is not explained by the regression line – it is due to random error) $(y_i - \hat{y})$.

Therefore:

$$(y_i - \bar{y}) = (\hat{y} - \bar{y}) + (y_i - \hat{y})$$

Total = Explained + Unexplained
Deviation Deviation Deviation

We can conduct these calculations for each and every value of Yi; to avoid negative figures, we square each calculation. When we sum them up, we obtain the following notation:-

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y_i - \hat{y})^2$$

Total Sum of Squares	=	Explained Sum of Squares	+	Unexplained Sum of Squares

Therefore: $\boxed{SST = SSR + SSE}$

Activity: In your own words define the following terms:

- Total variation
- Explained variation
- Unexplained variation

Further Formulae:

The Total Sum of squares is also given by:

$SST = (y_i - \bar{y})^2$; Which is also given by the formula:- $\sum y_i^2 - \frac{(\sum y_i)^2}{n}$

Alternatively: - SST (also denoted as Syy) = $\sum y_i^2 - n.\bar{y}^2$

The Regression Sum of squares is given by:

$SSR = (\hat{y} - \bar{y})^2$; Which is also given by the formula:-

SSR = $\beta_1^2 \sum (x_i - \bar{x})^2$

Therefore: SSR = $\beta_1^2[\sum x_i - \frac{(\sum xi)^2}{N}]$

Alternatively, the Regression (Explained) Sum of Squares (Also denoted as Sxy) is given by the term:

$$SSR = \frac{S^2 xy}{Sxx}$$

And: $Sxy = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum xy - n.\bar{x}\bar{y}$

$$Sxx = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n.\bar{x}^2$$

Therefore: $SSR = \frac{\sum (xy - n.xy)^2}{\sum (x^2 - n.\bar{x}^2)}$

Then SSE (Residual or Error Sum of Squares) is found by subtracting: **SSE = SST – SSR**

Calculating the Coefficient of Determination (R^2): Intuitively, if a regression equation does a good job describing the relationship between two variables, the estimated Explained Sum of Squares (Sum of Squares due to Regression – SSR) should constitute a larger proportion of the Total Sum of Squares than the Error Sum of Squares (SSE). This proportion contributes by SSR can be computed by the ratio:

R^2 = SSR / SST

Therefore:

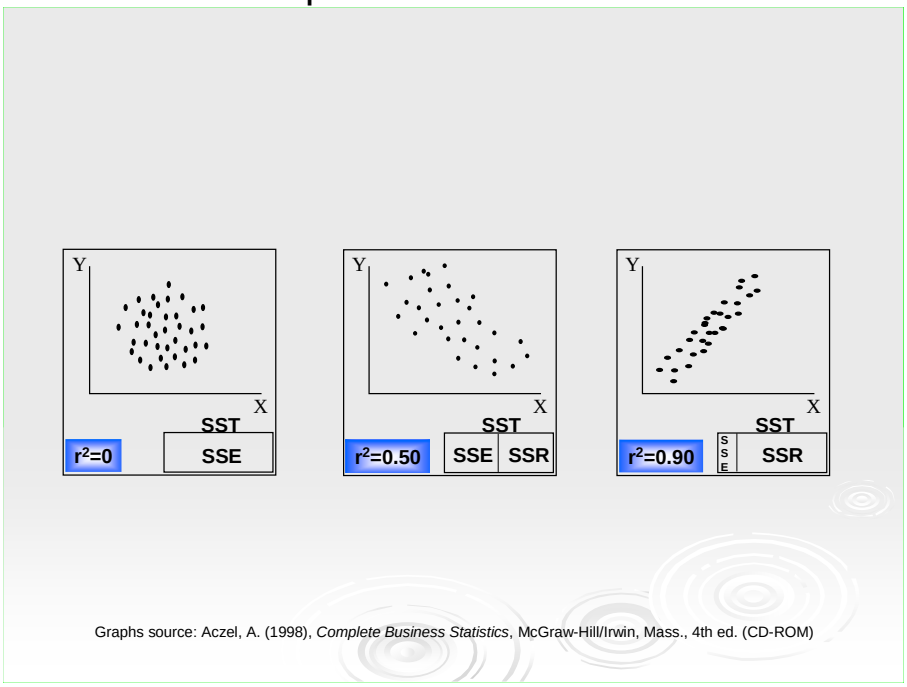
$$R^2 = \frac{\sum (y_i - \bar{y})^2}{\sum (\hat{y} - \bar{y})^2} = \frac{\beta_1^2[\sum x_i - \frac{(\sum xi)^2}{N}]}{\sum y_i^2 - \frac{(\sum y_i)^2}{N}}$$

The Coefficient of Determination, R^2 = **SSR / SST = 1 – SSE / SST**

Interpretation and use: The sample coefficient of determination measures the closeness of fit of the sample regression equation to the observed values of Y. When $(y_i - \hat{y})$ quantities (vertical distances from observed values of y from the estimated value on the line of best fit) are on average small, then the Un-explained Sum of Squares (Error Sum of Squares) is small and therefore a larger Explained or Regression Sum of Squares) and a large value of R^2. The largest value of R^2 is 1 and the lowest is 0. When R^2 is large it implies that the regression has accounted for a large proportion of the total variability in the observed values of y. Similarly when R^2 is low it implies that the regression has failed to account for a large proportion of the total variability in the observed values of y. It ought to be noted that we

cannot pass a final judgement on the model/equation until it has been subjected to another objective statistical test.

Illustration: Interpretation of R^2



Graphs source: Aczel, A. (1998), *Complete Business Statistics*, McGraw-Hill/Irwin, Mass., 4th ed. (CD-ROM)

We will later look at the relationship between coefficient of determination and the correlation coefficient.

ACTIVITY: Follow example 9.4.1 in Daniel page 420 in order to see the application of R^2

2. Testing H_0: $\beta_1 = 0$ with the F Test – Analysis of Variance for Linear Regression

The second method we can use to validate our linear equation is Analysis of Variance.

Recap of ANOVA: You will recall that in the method of ANOVA, the total variation present in s set of data is partitioned into two or more components. Associated with each of these components is a specific source of variation. In linear regression, we can observe that there are three types of variation: - The Total variation, the Explained variation and the Unexplained variation. We also indicated that we can square these measurements to obtain the Total Sum of Squares (SST), the Explained or Regression Sum of Squares (SSR), the Unexplained or Error Sum of Squares (SSE).

Remember that:

SST = SSR + SSE

The Total Sum of squares (Syy) is given by:

$$SST = (y_i - \bar{y})^2 = \sum y_i^2 - n.\bar{y}^2 \quad \text{..... (1)}$$

The Regression Sum of squares (Sxy) is given by:

$$SSR = (\hat{y} - \bar{y})^2 = \frac{S^2 xy}{Sxx} = \frac{\sum (xy - n.\bar{x}\bar{y})^2}{\sum (x^2 - n.\bar{x}^2)} \quad \text{..... (2)}$$

Therefore SSE can be found by subtracting: **SSE = SST – SSR**…….. (3)

These parameters can be summarised in an ANOVA Table as follows:

Source of Variation	**Sum of Squares**	**Df**	**Mean Square**	**Variance Ratio**
Linear Regression	SSR= $\frac{\sum(xy - n.\bar{x}\bar{y})^2}{\sum(x^2 - n.\bar{x}^2)}$	1	SSR/1	$\frac{MSR}{MSE}$
Residues (Errors)	SSE=SST – SSR	(n – 2)	SSE/(n – 2)	
	SST= $\sum y_i^2 - n.\bar{y}^2$	(n – 1)		

In general, the degrees of freedom associated with the SSR are determined by the number of constants minus one. In this case the constants are β_0 (the intercept) and β_1 (the slope); those of the SSE are n – number of variables (n – 2).

The test statistic is the Variance Ratio i.e.

$$\frac{MSR}{MSE} = VR$$

Distribution of the test statistic: When the hypothesis of no linear relationship between X and Y is true, and assumptions here-in are met, the Regression Mean Square (MSR) and the Residual Mean Square (MSR/MSE) is distributed as F with 1 and n – 2 degrees of freedom.

Decision Rule: Reject the null hypothesis H_0 if the computed value of the Variance Ratio (VR) is equal to or greater than the critical or tabulated value of F.

Calculation of the test statistic: We then calculate the test statistic and fill the appropriate calculations into the table.

Statistical Decision: We state the statistical decision in the following fashion: - Since the computed value is greater that the tabulated value at 1 and n – 2 df, the null hypothesis that $\beta_1=0$ is rejected.

Conclusion: If we reject the null hypothesis, them we conclude that the linear model provides a good fit to the data.

NB: ANOVA is a better estimator of than linear regression and correlation because it goes beyond just showing that an association is linear or not linear; It simply informs you that there is an association, whether linear, curvilinear, quadratic, polynomial, exponential, logarithmic etc, so that even if there is a relationship, but the relationship is not linear, ANOVA will demonstrate its existence.

Example: WORKED EXAMPLE ON ANOVA FOR SIMPE LINEAR REGRESSION

Anthropometric Data for 10 children is given in the table below:

Observation	Age	Weight
1	10	11
2	7	10
3	10	12
4	5	6
5	8	10
6	8	7
7	6	9
8	7	10
9	9	11
10	10	10

By the method of ANOVA, test whether the age of a child under 1 year explains their weight (test at a level of significance of 5%.

Answer: We first calculate the different measures and tabulate them as follows:

Observation	x	y	x^2	y^2	xy
1	10	11	100	121	110
2	7	10	49	100	70
3	10	12	100	144	120
4	5	6	25	36	30
5	8	10	64	100	80
6	8	7	64	49	56
7	6	9	36	81	54
8	7	10	49	100	70
9	9	11	81	121	99
10	10	10	100	100	100
$n=10$	$\sum x = 80$	$\sum y = 96$	$\sum x^2 = 668$	$\sum y^2 = 952$	$\sum xy = 789$

$$SSR = (\hat{y} - \bar{y})^2 = \frac{S^2 xy}{Sxx} = \frac{\sum (xy - n.\bar{x}\bar{y})^2}{\sum (x^2 - n.\bar{x}^2)}$$

Remember:

$$\bar{x} = \frac{1}{n_x}\sum x = \frac{80}{10} = 8$$

$$\bar{y} = \frac{1}{n_y}\sum y = \frac{96}{10} = 9.6$$

Substituting: $$[\frac{789 - 10(8)(9.6)}{688 - 10(8)^2}]^2$$

$$= 15.75$$

$$SST = (y_i - \bar{y})^2 = \sum y_i^2 - n.\bar{y}^2$$

Substituting: $952 - 10(9.6)^2 = 30.4$

$SSE = SST - SSR; 30.4 - 15.75 = 14.65$

Source of Variation	Sum of Squares	Df	Mean Square	Variance Ratio
Linear Regression	SSR=15.75	1	15.75/1	MSR=$15.75/1.84 = 8.56$ MSE
Residues (Errors)	SSE=14.5	(n – 2)	=14.65/(10 – 2) =14.65/8 =1.84	
	SST=30.40	(n – 1)		

From the tables, the tabulated value of F, $F_{0.05,df1,8} = 5.3177$. Since the computed value (8.56) is greater than the tabulated value, then at 5% level of significance, the available data set suggests that the weight of a child is significantly explained by the child's weight.

3. Testing H_0: $\beta_1 = 0$ with the t - test

If the assumptions stated earlier are true, then β_0 (the sample Y – Intercept) and β_1 (the sample slope) are unbiased point estimators of the Population Parameters (B_0 and B_1). Since under the assumptions, the sub-populations of Y are normally distributed, we may construct confidence intervals for and test hypotheses about β_0 and β_1. If the assumptions hold true, then the sampling distributions of β_0 and β_1 are normally distributed, with means and variances that approximate the population parameters.

Hypotheses regarding the Intercept β_0 are not normally of interest. It is the inferences regarding the slope β_1 that are of interest in testing hypotheses about the statistical significance of the linear relationship between X and Y.

The Test statistic: In this case, we have to "standardise" the value of β_1 before proceeding to test the hypotheses:

If the Population Standard deviation was known, then we would apply the Z – test; in this case, the Z – statistic would be given by:-

$$Z = \frac{b - \beta_1}{\sigma_b}$$

However, as a rule, the population standard deviation is not known, we therefore have to apply the t – test, in which case the t – statistic would be given by:-

$$t = \frac{b - \beta_1}{S_b}, df_{(n-2)}$$

The t – statistics follow a students' t-distribution with degrees of freedom (n – 2)

NB: To obtain S_b we must first estimate $\sigma^2_{y|x}$

$$\sigma_s = \frac{\sigma_{Population}}{\sqrt{n}}$$

Therefore:

$$S^2_{y|x} = \frac{\sum (y_i - \hat{y})^2}{n-2}$$

Conclusively:

$$S^2_{y|x} \text{ or } S_b = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

(The terms $(y_i - \hat{y})$ are called residues or error terms. Computer programmes for regression analysis routinely give these residues as part of the output. If this is the case, $S^2_{y|x}$ can be obtained by squaring all the residues and dividing the sum by $(n-2)$.

It follows then that since

$$S^2_{y|x} = \frac{\sum (y_i - \hat{y})^2}{n-2}$$

Then:

$$S^2_{y|x} = \frac{SSE}{n-2}$$

If we do not have the residues or error terms, the alternative formula is

$$S^2_{y|x} = \frac{n-1}{n-2}(S_y^2 - b^2 S_x^2)$$

Where: $S_y^2 =$ Variance of the y values

$S_x^2 =$ Variance of the x values

Steps:

1. Display and summarise the data
2. ***Assumptions:*** As for regression
3. Hypothesis: The hypotheses are:-
    a. H0: $\beta_1=0$
    b. HA: $\beta_1 \neq 0$

4. The test statistic: Given by:-

$$t = \frac{b - \beta_0}{S_b}$$

5. ***Calculation of the test statistic:***
    a. Remember that S_b is calculated by first finding $S^2_{y|x}$ using either the residues method or the variances of the y and x values.
    b. From $S^2_{y|x}$ then estimate S_b^2 and S_b
    c. Substitute into the equation above and calculate the t-statistic at degrees of freedom $(n-2)$. [Just for your memory – remember from Applied Biostatistics I that the equation for the test statistic above is called the "Critical ratio".

6. ***Statistical decision:*** Reject Null hypothesis if the calculated t is in the rejection zone. [e.g., if $n = 30$, then the critical value of t, at df $(30-1)$ and $\alpha = 0.05$ is 1.2896. Therefore, we reject H0 if the computed value of t is greater than or equal to – 1.2896 for a two tailed t – test.
7. ***Conclusion:*** If the t – statistic is out of the rejection zone, then we conclude that the slope of the line is not equal to zero, and that there is indeed a linear relationship of some sorts. This implies in practice that we can expect to get better predictions and estimates of Y if we use the sample regression equation that if we ignore the relationship between X and Y. This leads us to the conclusion that based on the sample findings, the relationship between X and Y in the population is also a direct linear relationship.

Confidence Interval for β_1 : Remember from Applied Biostatistics 1 that the formula for confidence interval is:

[(Estimator) $\pm$ (Reliability Factor) (Standard Error of the Estimate)]

Therefore, the CI for β_1 is:

$$b \pm t_{(1-\alpha/2)} \cdot \sqrt{\frac{S^2_{y|x}}{\sum (x_i - \bar{x})^2}}$$

Or:

$$b \pm t_{(1-\alpha/2)} \cdot \sqrt{\frac{S^2_{y|x}}{\sum x^2_i - (\sum x_i)^2 / n}}$$

Statistical Decision: Accept H_0 if CI includes zero; reject it if the CI does not include zero because $\beta_0=0$ cannot be a possible figure in a CI that does not include zero.

g. Step 4: Using the Regression Equation

If the data conform satisfactorily to a linear model, use the equation to **Predict** and to **Estimate.**

Prediction: If we know one parameter Y for a given value of X e.g. we can estimate the expected weight of a child if we know their age or height.

Estimation: We can estimate the mean of a variable Y for a given population for a given value of X e.g., we can estimate the mean Weight of class of pupils if we know their mean age.

NB: Failure to reject H_0 does not mean that X and Y are not related; there could be some other explanation of their relationship that is not linear. It may be curvilinear, logarithmic, exponential etc. On the other hand, accepting H_0 does not also mean straight away that the best explanation for the relationship is that it is linear – rather, it may indeed be curvilinear or logarithmic, but some linearity provides a good fit. This is one of the set-backs of linear regression.

Session 2: Correlation Analysis

Introduction: In linear regression, we examined the relationship of a fixed variable X to a continuous variable Y. We can use the method of correlation for two ends:

1. We can analyse further the findings of a linear regression, in which we find the conclusion of there being a linear relationship between X and Y, purposefully to determine the strength of association.
2. We can also use correlation to analyse the relationship between two variables are both continuous in nature (none has fixed values) (Sometimes it may not be clear cut as to which is the independent or dependent variable). These characteristics are called correlates.

Illustration: We know that Height and weight in children are both continuous variables that are not fixed (there are no fixed heights in a population, but rather a continuum of heights); the same applies to weight. We can then determine the strength of the association between height and weight.

Therefore correlation is a measure of that shows how two random variables are associated in a sample. We usually use **r** to indicate a sample correlation coefficient.

Session Topics: The following topics will be covered:

a. Types of correlation coefficients
b. Correlation Assumptions
c. The Correlation Coefficient
d. Formulae for r and r^2
e. Testing Hypotheses about the value of r :

Session Objectives:

By the end of this Session, the MPHO should be able to:

1. Evaluate datasets to discriminate the variables that can lend themselves to a correlation analysis
2. State and evaluate the assumptions in Correlation
3. Describe the nature and characteristics of the correlation coefficient
4. Calculate the correlation coefficient with a given dataset of two correlates
5. Formulate and test hypotheses about the value of the correlation coefficient

a. Types of correlations

Types of correlation include:

Pearson correlation: This is a parametric method which assumes bivariate normality.
Spearman rank correlation: Non parametric i.e., no distributional assumption.

b. Correlation Assumptions

- For each value of X there is a normally distributed subpopulation of Y values
- For each value of Y there is a normally distributed subpopulation of X values
- The joint distribution of X and Y is normally distributed (bivariate normal distribution)
- The subpopulation of Y have the same variance
- The subpopulation of X have the same variance

c. The Correlation Coefficient

The bi-variate Normal Distribution above has 5 parameters:

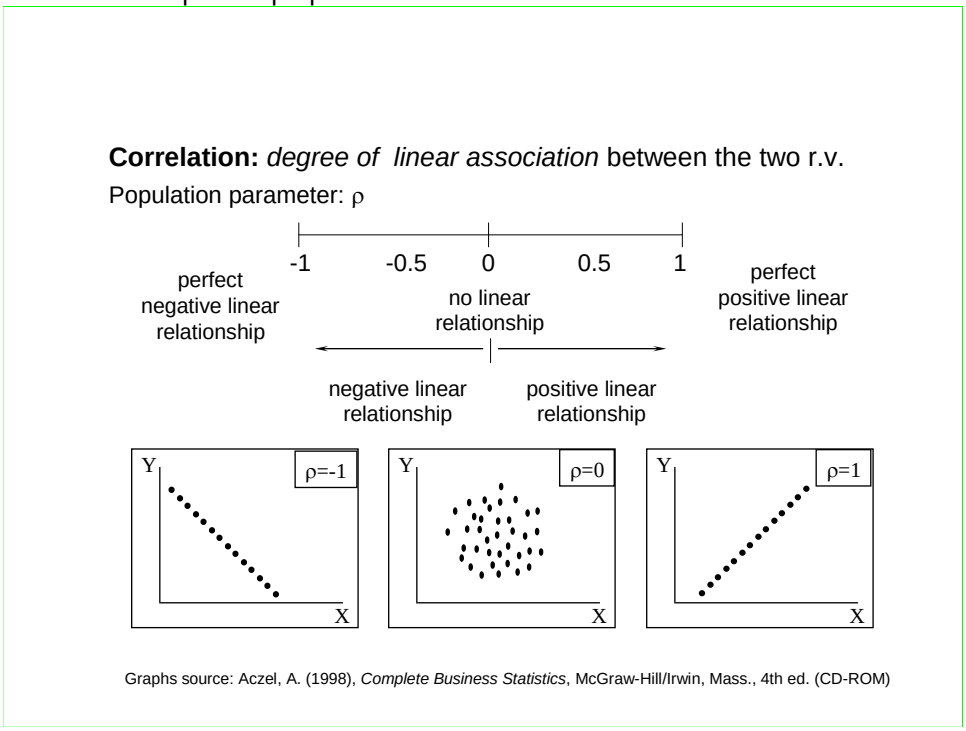
1. σ_y = Standard deviation for distribution y
2. σ_x = Standard deviation for distribution x
3. μ_y = Mean for distribution y
4. μ_x = Mean for distribution x
5. ρ = The Population correlation coefficient – it measures the strength of the linear relationship between X and Y.

ρ is the square root of ρ^2, the Population Coefficient of Determination. It is approximated by the sample correlation coefficient r which is indeed the square root of the sample coefficient of determination r^2. Since the Coefficient of Determination (r^2) takes on values that range between 0 and 1, then the Correlation Co-efficient (its Square root) may range between – 1 to + 1.

If r = 1; then we have perfect direct linear correlation between X and Y

If r =– 1; then we have perfect direct linear correlation between X and Y

Illustration: Important properties of the correlation coefficient are summarized below:



Graphs source: Aczel, A. (1998), *Complete Business Statistics*, McGraw-Hill/Irwin, Mass., 4th ed. (CD-ROM)

d. Formulae for r and r^2

Since: $$r^2 = \frac{\beta_1^2[\sum x_i - \frac{(\sum xi)^2}{N}]}{\sum y_i^2 - \frac{(\sum y_i)^2}{N}}$$

Then: $r = \sqrt{\dfrac{\beta_1^2[\sum x_i - \dfrac{(\sum xi)^2}{N}]}{\sum y_i^2 - \dfrac{(\sum y_i)^2}{N}}}$

e. Testing Hypotheses about the value of r ***:***

1. Examine the data as given, to determine if there is possible correlation between the two variables
2. ***Assumptions:*** The Assumptions of Correlation hold
3. ***Hypotheses:*** $H_0 : \rho = 0; H_A : \rho \neq 0$
4. ***The test statistic:***

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

5. ***Distribution of the test statistic:*** It follows a student's t – distribution with n – 2 degrees of freedom.
6. ***Decision rule:*** If the calculated test statistic is greater than the corresponding critical value of t at degrees of freedom (n – 2), then reject H_0
7. ***Statistical Decision:*** If the calculated H_0 exceeds the critical value of t , then reject H_0.
8. ***Conclusion:*** If H0 is rejected, then we conclude that population X and Y are linearly correlated.

3.0 ADDITIONAL RESOURCES

3.1 TEXT DOCUMENTS FOR ADDITIONAL READING

Text Document 3.1.1:

Summary of important formulae

Important Parameters

Mean = $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Variance = $\frac{n\Sigma x^2 - (\Sigma x)^2}{n(n-1)}$

SD for proportions: $s_p = \sqrt{\frac{p.(1-p)}{n}}$

The variance of the difference between proportions:

$\sqrt{\frac{p_1.(1-p_1)}{n_1}} + \sqrt{\frac{p_2.(1-p_2)}{n_2}}$

Variance of difference between means

$\sigma_{(x_1-x_2)} = \frac{\sigma^2}{n} + \frac{\sigma^2}{n}$

Probability:

Exhaustive events: For one set of events E_n:
$(PE_1) + (PE_2) + (PE_3) + \ldots.. + (PE_n) = 1$
Mutually exclusive events:
$P(A \cup B) = P(A) + P(B)$
Multiplication Rule:
Joint probability = Conditional probability X Marginal probability
i.e. $P(M \cap X) = P(X \mid M) \times P(M)$

Independent events:
$P(A \cap B) = P(A) \times P(B)$
Complementary events: $P(A \cap B) = P(A) + P(B) = 1$
The Addition Rule:
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Baye's Theorem:

$$P(D \mid T) = \frac{P(D \cap T)}{P(T)} = \frac{P(T \mid D) \cdot P(D)}{P(T \mid D)\,P(D) + P(T \mid D^C)\,P(D^C)}$$

Total Probability:
Assuming $A_1 \ldots A_K$ are mutually exclusive/exhaustive events,

Then:
$P(B) = P(B \mid A_1).\ P(A_1) + P(B \mid A_2).\ P(A_2) + \ldots\ldots\ldots\ldots + P(B \mid A_K).\ P(A_K)$
Therefore: $\sum_{i-1}^{K} P(B \mid A_i) \cdot P(A_i)$

Relationships of estimators to population parameters

	Population	Population parameter	Sample	Estimator
1.	Population mean	(μ)	Sample mean	$(\bar{x})$
2.	Population SD	(σ)	Sample SD	(s)

3.	Difference between population means	$(\mu_1 - \mu_2)$	Difference between sample means	$(\bar{x}_1 - \bar{x}_2)$ $se = (\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$
4.	Population proportion	(P)	Sample proportion	(p)
5.	Difference between population proportions	$(P_1 - P_2)$	Difference between sample proportions	$(p_1 - p_2)$ $se = \sqrt{\frac{p_1.(1-p_1)}{n_1}} + \sqrt{\frac{p_2.(1-p_2)}{n_2}}$

Z and t statistics

Z- Statistic for population means (population SD known):

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma}}$$

Z-statistic for mean of samples (population SD known):

$$z = \frac{\bar{x}_i - \mu}{\sigma / \sqrt{n}}$$

Z – Statistic for the sample proportion

$$z \equiv \frac{p - P}{\sqrt{\frac{p.(1-p)}{n}}}$$

t – Statistic for mean (Population SD unknown (at $df\,(n-1)$) :

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Confidence Intervals

Confidence Interval for mean; Population SD known:

$[\bar{x} \pm 1.96 * se]$ (For 95% CI)

Or

$[\bar{x} \pm z_{(\alpha/2)} * se]$ (For $(1-\alpha)\%$ CI)

or

$[\bar{x} \pm t_{\alpha, df\,(n-1)} * s / \sqrt{n}]$;(For $(1-\alpha)\%$ CI at a sample size of n)

Confidence interval for mean; Population SD not known:

$[\bar{x} \pm 2.042, _{df\,(30)} * s / \sqrt{n}]$;(For 95% CI and a sample of size 31)

Confidence Interval for proportions

$$p \pm z_\alpha * \sqrt{\frac{p.(1-p)}{n}}$$

Dealing with two independent samples

Sample characteristics	
Sample 1	**Sample 2**
n_1	n_2
Sample mean 1 = $\bar{x}_1$	Sample mean 2 = $\bar{x}_2$

Variance 1 = $s_1^2 = \frac{\sum(x_i - \bar{x}_1)}{n_1 - 1}$	Variance 1 = $s_2^2 = \frac{\sum(x_i - \bar{x}_2)}{n_{21} - 1}$
Pooled estimate of variance	
$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$	
But remember: The standard deviation for this relationship was: $\frac{\sigma}{\sqrt{n_1}} + \frac{\sigma}{\sqrt{n_2}}$ In this case, we use the sample standard deviation since we do not know the population one. Thus, $se = \frac{s_p}{\sqrt{n_1}} + \frac{s_p}{\sqrt{n_2}} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	
Test statistic	
It has a t-distribution with degrees of freedom: $(n_1 + n_2 - 2)$ $t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	

Sample Size

Sample size for mean without power:

$$n = \frac{z^2.\sigma^2}{Error^2}$$

Sample size for proportion without power:

$$n = \frac{z^2.p(1-p)}{Error^2}$$

Sample size with power (mean)

$$n = \frac{(z_1 + z_2).\sigma^2}{Error^2}$$

Where Error = $\mu_0 - \mu_1$

Sample size with power (proportion):

$$n = \frac{(z_1 + z_2).p(1-p)}{Error^2}$$

Where Error = $\mu_0 - \mu_1$

Confidence Intervals for RR and OR

The SE of the natural log of the RR (In RR) is given by:

$$se * InRR = \sqrt{\frac{1}{a} - \frac{1}{(a+b)} + \frac{1}{c} + \frac{1}{(c+d)}}$$

Therefore, the 95% CI for In RR is:

L1 In (RR) – 1.96 X SE (In RR)

L2 In (RR) + 1.96 X SE (In RR)

The CI for the RR is thus the antilog e_1^L; e_2^L

The SE for the natural log of the OR (In OR) is:

$$se * InOR = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Therefore, the 95% CI for In RR is:

L1 In (OR) – 1.96 X SE (In OR)

L2 In (OR) + 1.96 X SE (In OR)

The CI for the OR is thus the antilog e_1^L; e_2^L

The Chi-square (X^2)

$$x^2 = \frac{(ad - bc)^2 N}{(a+b)(c+d)(a+b)(b+d)}$$

df = (Columns-1) + (Rows-1)
i.e. (2 – 1) + (2 – 1) = 2

3.2 GLOSSARY OF TERMS

Computer: is a general-purpose device with a capacity to input data, store it, process it and output it in the required form. That is, a Computer was designed to handle instructions as well as data to be processed

Continuous variables: A continuous variable can assume any value within a specified relevant interval of values assumed by a variable. A continuous random variable does not possess gaps or interruptions like the discrete variable. Examples of continuous variables include, time, blood pressure and weight.

Data: Data is a raw material of statistics. In statistical terms, we define data as a number. You will find out that in statistics, we use two kinds of numbers; those that result from counting i.e. number of patients being discharged from a hospital and another form are those numbers that result from measurements such as taking a patient's temperature or a nurse measuring a child's weight.

Descriptive Statistics: The aspects of organization, presentation and summarizing of data are labelled as descriptive Statistics. This was the main subject matter of statistics up to the 20th century. In health science, descriptive statistics are used to characterize the health status of a particular population. Examples of descriptive statistics include measures of location and measures of central tendencies that will be explained later.

Discrete Random variables: Such a variable is always characterized by gaps or interruptions in the values that it can assume. Such gaps indicate absence of value between particular points.

Hypothesis: It is a 'statement', 'idea', or an 'allegation' made about the true parameters of the population.

Inferential Statistics: On the other hand, statistical inference is concerned with the logical basis by which conclusions regarding populations are drawn from results obtained from a SAMPLE. Hypothesis testing is an example of inferential statistics.

Interval estimation: The method whereby the true value of the parameter is given between limits with a certain level of confidence. The confidence interval can also be looked at as a measure of precision.

Point estimation: The statistical method (technique) which when applied gives a single value of the estimator, e.g. the mean.

Population: A population is defined as a collection of all elements for which we have an interest at a particular time point. If you take a measurement of some variables on each entry in a population, we generate a population of that value of that variable. Some examples of populations are the population of vehicles in Kampala, population of all medical school students and the population of children under five years who are malnourished.

Probability: The degree of certainty or uncertainty of occurrence of an event.

Qualitative variables: These are characteristics that are not capable of being measured. Such characteristics are categorized for example; skin colour, behaviour etc. Qualitative variables always convey information regarding attributes.

Quantitative variables: These are variables that can be measured in the usual sense or numerical form. Such measurements normally convey information on the amount. For example; number of students admitted in a year,

Random variable: When values of particular variables are obtained as a result of chance factors such that they cannot be predicted in advance such a variable is referred to as a random variable.

Variable: In the subject of statistics, you will find that a variable is defined as a characteristic or attribute that takes on different values in different persons, places, things or time. Such a characteristic is not the same when observed in different possessors.

3.3 REFERENCES

Wayne, W.D. (1998): Biostatistics: A Foundation for Analysis in the Health Sciences. John Wiley & Sons, Inc. 7th Edition

Bernard, R. (19##): Fundamentals of Biostatistics, 5th Edition

Brand, M. (1987): An introduction to medical statistics; Oxford University Press, Oxford.

Altman, D. (1991): Practical statistics for medical research; Chapman & Hall, London.

Lwanga, S. (1975): Biostatistics for medical students; East African Publishers, Nairobi

Kirkwood BR, Sterne JAC (2003): Essential Medical Statistics, Blackwell Publishing Company, 2nd Edition

3.4 ANSWERS TO QUIZ QUESTIONS

ANSWERS TO QUIZ 2.1.1

1.d 2.c 3.d 4.a

ANSWERS TO QUIZ 2.2.1

1.d 2.a 3.c 4.b 5.b 6.a 7.d 8.c 9.c 10.a

ANSWERS TO QUIZ 2.3.1

A1. When we standardize 24: = - 2.5
When we standardize 25: = - 1.25
The corresponding p-values are 0.1056 to 0.0048
Therefore, the probability is the difference 0.1056 – 0.0048 = 0/1008

A2. The difference between the given means is 0.9
The variance of this difference is 0.05; therefore, the standard deviation of this difference is 0.02236.
When we standardize our hypothesized difference of 1, we get 0.45
Therefore P ($(x_1 - x_2)$ > 1.0) which is the P (z > 0.45) = 0.3264

A3. The mean = 0.55; variance = (0.55X 0.45)/200; therefore, SD = 0.0352
Standardizing 0.45: (0.45 – 0.55)/0.0352 = - 2.84
This corresponds to: P ($z \leq$ - 2.84) = 0.0023

A4 The formula is: $[X_S - Z\alpha/_2 \,.\, (\sigma_X)/\sqrt{n} < \mu < X_S + Z\alpha/_2 \,.\, (\sigma_X)/\sqrt{n}]$

Therefore: $7 - 1.96 \,.\, (3)/\sqrt{25} > \mu > 7 + 1.96 \,.\, (3)/\sqrt{25}]$
5.82 < **μ** < + 41.96; The 95% CI is therefore: **5.82** to **8.18**

A5. The formula is: $[X_S - t_{\alpha,\, df\,(n-1)} \,.\, (\sigma_S)/\sqrt{n} < \mu_X < X_S + t_{\alpha,\, df\,(n-1)} \,.\, (\sigma_S)/\sqrt{n}]$
Substituting: $[68.7 - 2.093 \,.\, (13.02)/\sqrt{20} < \mu_X < 68.7 + 2.093 \,.\, (13.02)/\sqrt{20}]$
[**62.61** < μ_X < **74.79**]; therefore, the 95% CI for the estimate is: 62.61 to 74.74

A6. The sample proportion is therefore: $\frac{80}{1000}$ = 0.08

Substituting: **0.08 – 1.96.** $\sqrt{\frac{0.08\ .(1-0.08)}{200}}$ **< P< 0.08 + 1.96.** $\sqrt{\frac{0.08\ .(1-0.08)}{200}}$

Therefore: 0.042 < P < 0.118; the 95% CI is: 0.042 to 0.118

A7. The 95% CI is given by $\bar{x}$ + (1.96 x $\sigma/\sqrt{n}$)

= 97.2 + 1.96(0.2)/ $\sqrt{100}$

= 97.2 + 1.96(0.2)/10

= 97.2 + 0.04

= 97.16, 97.24

A8. You should obtain the following confidence intervals [97.08, 97.32]; they are wider than the earlier ones. If we use a smaller sample, we get a less precise estimate.

A9. You should obtain the following confidence intervals [97.12 to 97.28]; they are slightly wider than the earlier ones, meaning that an increase in standard deviation increases the confidence intervals

A10. H_0: $\mu = 268$; H_A: $\mu \neq 268$

Test to be done: Z – test; two tailed

Test statistic: $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

$= \frac{272.5 - 268}{15/\sqrt{25}} = 1.5$

Critical values: At α= 0.05, the critical values are $\pm$ 1.96

Decision: Since 1.5 is not beyond 1.96; we do not reject the null hypothesis

Conclusion: There is no evidence that an average pregnancy term is not 268 days

P-value: The corresponding p-value for a Z score of 1.5 **from the z-table** is 0.1336; since this is greater than 0.05, it is in the non-rejection zone. We therefore do not reject the null hypothesis.

A11. H_0: $\mu \leq 268$; H_A: $\mu > 268$

Test to be done: Z-test (one tailed)

Test statistic: $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

$= \frac{272.5 - 268}{15/\sqrt{25}} = 1.5$

Critical values: At α= 0.05, the critical value is $\pm$ 1.645 (using one tailed z tables).

Decision: Since 1.5 is not beyond 1.645; we do not reject the null hypothesis

Conclusion: There is no evidence that the average period of stock-out of all three indicator essential drugs is more than 268 days; evidence points to the fact that it is at most 268 days.

P-value: The corresponding p-value for a Z score of 1.5 **from the z-table** is 0.0668 for a **one tailed** table, corresponding to a cut off of α = 0.05 for only this region. Since the p-value of 0.06 this is greater than 0.05, it is in the non-rejection zone. We therefore do not reject the null hypothesis.

A12. H_0: $\mu = 268$; H_A: $\mu \neq 268$

Test statistic: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

$$= \frac{272.5 - 268}{12/\sqrt{36}} = 2.25$$

The degrees of freedom are: (n-1) = 36-1 = 35

Critical values: At α= 0.05, and df 35, the critical values are $\pm$ 2.0301 (using two tailed z tables)
Decision: Since 2.25 is not beyond 1.96; we reject the null hypothesis
Conclusion: There is evidence that an average pregnancy term is not 268 days
P-value: The corresponding p-value for a t - score of 2.25 **from the two tailed t-table** is between 0.02 and 0.05. Since this is less than 0.05, it is in the rejection zone. We therefore reject the null hypothesis.

A13: H_0: $\mu_X \geq 140$
H_A: $\mu_X < 140$
Test to be done: t – test (one tailed)
Test statistic: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

$$= \frac{138.47 - 140}{2.66/\sqrt{20}} = -2.57$$

The degrees of freedom are: *(n-1)* = 20-1 = 19

Critical values: At α= 0.05, and df 19, the critical values are – 1.7291 (using one tailed z tables)
Decision: Since – 2.25 is not beyond – 2.57 we reject the null hypothesis
Conclusion: There is evidence that the average capacity of theatre lamp inverters in the general population is not more than 140 ampere – hours. It is in fact less than 140 ampere – hours.
P-value: The corresponding p-value for a t - score of -2.57 **from the one tailed t-table** is even **less than 0.01**. Since this is less than 0.05, it is in the rejection zone. We therefore reject the null hypothesis.

A14. H_0: $p \geq 0.10$; H_A: $p < 0.10$
Test to be used: Z – test for proportions (one tailed)
Test statistic: $z = \sqrt{\frac{p(1-p)}{n}}$

$$= \frac{(11/200) - (10/100)}{\sqrt{\frac{0.10.(1-0.10)}{200}}}$$

Critical values: At α= 0.05, the critical values are – 1.645 (using one tailed z tables)
Decision: Since – 2.12 is beyond – 1.645, we reject the null hypothesis
Conclusion: We reject the null hypothesis and conclude that there is evidence that the new packaging system is less than 10% defective.
P-value: The corresponding p-value for a Z score of – 2.12 **from the z-table** is 0.0170 for a **one tailed** table.

A15. (1) $X^2 = \frac{(56X\ 1927 - 84X552)^2}{607X2711X139X2479}$

X^2 = 22.122, df =1 and p<0.001. We reject the hypothesis that probability of diabetes is the same for both cases and controls.

(2) Estimated Odds Ratio (OR) $= \frac{55X1927}{84X552} = 2.286$

(3) The 95% confidence limits for ln (OR) are given by

(4)Therefore, the 95% CI for the OR is ($e^{0.827-1.96 X 0.180}$, $e^{0.827+1.96 X 0.180}$) = (1.61, 4.57)

3.5 INDEX OF URLs FOR INTERNET RESOURCES

1. http://encyclopedia.thefreedictionary.com/Descriptive%20statistics
2. http://www.itc.virginia.edu/desktop/docs/fms/pc/organize.html
3. http://www.phppo.cdc.gov/phtn/catalog/pdf-file/LESSON3.pdf

4. http://www.cdc.gov/descd/MiniModules/PPS/page01.htm : FYI, use and dissemination a 30 minute web training on rationale and methods for PPS (probability proportionate to size) sampling

5. http://www.training.nih.gov/careers/careercenter/publish.html : Improve your writing skills

6. http://www.idrc.ca/IMAGES/books/WFC_English/WFC_English/sitemap.html: Improve your writing skills
7. http://www.son.wisc.edu/rdsu/index.html: This is the SPSS Web site. You can access tutorials on SPSS on this site. On the site, go to the link: **LIBRARY**. You can access several tutorials. There is a real-time voice attachment to the presentations – you can follow the lessons and learn by yourself. Hope you will enjoy the tutorials.

3.6 INDEX OF DISCUSSION FORUM QUESTIONS

You are encouraged to participate in the discussion forums that have been pre-planned for you in the semester. Through these forums, you will be able to exchange information with the moderators and fellow students and gain a deeper understanding of the material you have read. There will be at-least one discussion forum for each course, with a number of questions drawn from each course unit. You are requested to post your discussion points to the board for other members to share. Please be brief and to the point. You may discuss only one question or a number of them depending on where you feel motivated. You may also post a discussion point that is outside the set questions, provided you have **HOT** points to share. These forums will enhance the "virtual" classroom environment and facilitate you to learn at the same pace as the others.

There is a detailed outline of the schedule of these forums that will be handed to you at the beginning of the semester, under the resource: **SEMESTER SCHEDULES**. At the precise times indicated (Modifications in the schedule may from time to time be communicated to you by the Moderator), the discussion will be activated and you will be called upon to contribute; this will be a "silent" online call – you are requested to remain alert, and regularly check the forum platform or your internet mail-box for the call. For each course the discussion will run for an entire week.

The forums will either be hosted at the Makerere University Intranet site **http://intranet.mak.ac.ug**, where you follow the link "**Forums**", or may be posted at the e-learning tool on the site

http://nextgen.mak.ac.ug. Please sign up and register your unique identity in the forum. The following is a summary of the questions that are up for discussion in the Discussion Forum for this particular Course:

1. *Statistical Methods make use of several measures of location and dispersion in describing data. Select one measure and describe its computation and application*

2. *Statistical Methods make use of several sampling distributions in testing hypotheses about population characteristics using sample information. Select one sampling distribution and briefly describe its characteristics, parameters and when it is used*

3. *There are diverse hypothesis testing scenarios that depend on the directionality of the hypothesis to be tested as well as whether certain population parameters are known. Select one scenario and illustrate, with a numerical example, its salient features*

POST YOUR REPLY NOW:

Post your reply **now** to one or more of these issues and attend the Forum; you will discover the unique learning experience from sharing knowledge in this interesting resource!

3.7 INDEX OF ADDITIONAL RESOURCES FOLDER

a. Introduction to the Statistical Package for Social Sciences
b. Introduction to EpiInfo
c. Introduction to other Internet resources
d. Statistical Tables
e. Introduction to E-Learning Platforms
f. Tables of random numbers
g. Basic statistical concepts and Univariate analysis

3.8 INDEX OF SELECTED REAL TIME LECTURE NOTES

1. Comparing Two Independent Samples
2. Descriptive Statistics
3. Estimation
4. Hypotheses Testing
5. Probability Concepts
6. Sample Size Determination
7. Sampling Distributions
8. Two by Two Contingency Tables
9. Basic Statistical Concepts
10. Data Management
11. Data presentation
12. Measures of location and dispersion

3.9 SUMMATIVE EVALUATION OF THE INSTRUCTION PROCESS

Summative Evaluation of the instruction will be conducted using the following means:

1. Progressive Assessment in form of Hand-in Assignments
2. Progressive Assessment Test – To be done during the Face – to – Face Sessions
3. The University Examination
4. An optional Post-test
5. A Course Post Evaluation Questionnaire

Progressive Assessment – Hand-in Assignments

These assignments should be handed in by the time of sitting for the progressive assessment test at the institute of public health. They will be marked and will contribute to the final progressive score. The number, nature and timing of assignments will be determined by the Course Coordinator. Some of these assignments may be directly included in these materials by the Course Coordinator. An index of them is listed below:

	TITLE	COMENTS
1.		
2.		
3.		
4.		
5.		

Post-test

It is optional for you to attempt this **Post-test**. It will assist you to gauge your grasp of the material after the instruction process. The test is contained in the **Additional Resources Folder**.

Post Evaluation

We are in need of your feedback on the quality and content of these materials. It will be valuable to the iterative process of their further improvement. For this purpose, we have attached a questionnaire to gauge your perception of the design and conduct of this course and to link this to your understanding of the subject matter. This questionnaire has been introduced to you previously. It contains two parts: a Post Evaluation of the materials for the previous semester and a pre-evaluation of the materials you expect for this semester. Please note that this is not a progressive assessment or exam, and will not contribute to your final mark. It should be completed at the beginning of the semester. Please fill in the Evaluation questionnaire for this semester; make comments as requested, and send it as an e-mail attachment or hardcopy to: **Dr. Roy William Mayega – Instructional Designer/Editor – MPH Distance education Programme:** e-mail: de_materials@musph.ac.ug .