NATIONAL OPEN UNIVERSITY OF NIGERIA

FACULTY OF HEALTH SCIENCES

DEPARTMENT OF PUBLIC HEALTH SCIENCE

CODE: PHS813

COURSE TITLE: BIOSTATISTICS AND APPLICATIONS

COURSE GUIDE

PHS813: BIOSTATISTICS AND APPLICATIONS

COURSE DEVELOPER/WRITERS: **Dr. Eno E. E. Akarawak**
Department of Mathematics
University of Lagos, Akoka

Dr. Rotimi K. Ogundeji
Department of Mathematics
University of Lagos, Akoka

COURSE EDITOR: **Dr. Matthew Olaniyi Olayiwola**
Department of Statistics
Federal University of Agriculture, Abeokuta

COURSE COORDINATOR: **Dr. Florence N. Uchendu**
Department of Public Health
National Open University of Nigeria

PROGRAMME LEADER: **Prof. Grace Okoli**
Faculty of Health Sciences
National Open University of Nigeria

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

Lagos Office:
14/16 Ahmadu Bello Way
Victorial Island, Lagos

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

Printed:

ISBN:

CONTENTS **PAGE**

INTRODUCTION

Biostatistics is a branch of study that concerns itself with the applications of statistical methodologies to a wide range of topics in biological situations. Essentially, it ensures that findings and practices in public health, biological and agricultural sciences and biomedicine are supported by reliable evidence. Statistics is central to providing such needed reliable evidence to support research findings.

PHS 813: Biostatistics is a first semester three-credit course for M.Sc. Public Health. This course guide covers topics in Biostatistics that will introduce students to statistical procedures relating to Bio sciences. The course guide will also expose students to statistical tools that can be applied in research on health-related fields, including Medicine, Epidemiology and General Public health. It is written to give students a better understanding of Biostatistics and its applications.

The course guide tells you briefly what the course is all about, what course materials you will be using, and how you can work your way through these materials. It suggests some general guidelines for the amount of time you are likely to spend on each unit of the course in order to complete it successfully. It also outlines some tutor-marked assignments. There may be regular tutorial classes that are linked to the course. You are advised to attend these sessions whenever they come up.

WHAT YOU WILL LEARN IN THIS COURSE

The overall aim of PHS 813: Biostatistics is to introduce some statistical methodologies in medical research; help students understand some of the statistical principles of good practice in medical investigations and how to use and interpret some of the statistical techniques used in Biostatistics.

COURSE OBJECTIVES

In order to achieve the aims, set out above, the course sets overall objectives. In addition, each unit also has specific objectives. The unit objectives are always included at the beginning of each unit and these should be read before you start working through the unit. You may need to refer to them during your study of the unit in order to check on your progress. You should always look at the unit objectives again after completing each unit. In this way, you can be sure that you have achieved what was required of you by the unit. Set out below are the wider objectives of the course as a whole. By meeting these objectives, you should have achieved the aims of the course as a whole. On successful completion of the course, you should be able to:

i. define and classify Statistics
ii. describe different methods of data collection and presentation used in medical research
iii. explain the different descriptive techniques used in obtaining summary statistics
iv. evaluate the probability of an event, both in experimental and theoretical situations
v. explain different sampling procedures and sampling distributions
vi. discuss different estimation procedures
vii. explain the various hypothesis testing procedures for decision making
viii. analyze bivariate data using correlation and regression
ix. apply the various statistical techniques in medical and non-medical situations

WORKING THROUGH THIS COURSE

To complete this course, you are required to read the study units and books, and other materials provided by the National Open University of Nigeria (NOUN). Each unit contains self-assessment exercises. At the end of the course, there is a final examination. The course should take you about 14 weeks to complete. Below, you will find listed all the components of the course, what you have to do, and how you should allocate your time to each unit in order to complete the course successfully and on time.

COURSEMATERIALS

Major components of the course are:

i. The Course Guide
ii. Study Units
iii. References
iv. The Presentation Schedule

STUDY UNITS

Each study unit consists of some work, and includes introduction, specific objectives, reading materials, conclusion, summary, tutor-marked assignments (TMAs), references and further readings. The units direct you to work on exercises related to the required readings. In general, these exercises are on the material you have just covered. Together with tutor-marked assignments, these exercises will assist you in achieving the stated learning objectives of the individual units and of the course.

THE ASSIGNMENT FILE

The course assignment will cover:

i. Definitions and basic concepts; data collection and presentation methods; data summary
ii. Descriptive statistics for ungrouped and grouped data
iii. Probability, random variables and their distributions
iv. Sampling and sampling distributions
v. Estimation: point estimation and confidence interval

vi. Hypothesis testing procedures: one sample, two samples, more than two samples, categorical data and bivariate data.

ASSESSMENT

There are two aspects to the assessment of the course:

i. Tutor-Marked Assignments,
ii. Written examination.

TUTOR-MARKED ASSIGNMENTS (TMAs)

There are some tutor-marked assignments. You are encouraged, however, to submit the assignments to be given to you at the Study Centre for this course. In tackling the assignments, you are expected to apply information, knowledge and strategies gathered from the relevant study units. However, it is desirable to demonstrate that you have read and researched more widely than the required minimum. Using other references will give you a broader viewpoint and may provide a deeper understanding of the subject.

The assignments must be submitted together with a TMA (Tutor-Marked Assignment) form, to your tutor for formal assessment in accordance with the deadlines stated in the Presentation Schedule and the Assignment File. If, for any reason, you cannot complete your work on time, contact your tutor before the assignment is due to discuss the possibility of an extension. Extensions will not be granted after the due date unless there are exceptional circumstances. The work you submit to your tutor for assessment will count for 30% of your course mark.

FINAL EXAMINATION AND GRADING

The final examination for PHS 813 will be for 2 hours and its result will contribute 70% to the total course grade. The examination will consist of questions which reflect the types of self-testing, practice exercises and tutor-marked problems you have previously

encountered. All areas of the course will be similarly assessed. Use the time between when the last unit was read and sitting for the examination to revise the entire course. You might find it useful to review yourself-assessment questions, tutor-marked assignments and comments on them before the examination.

COURSE MARKING SCHEME

The following lays out how the actual course marking is broken down and summarized in Table 1.

Table 1: Assessment Marks

Assignments	**Marks**
Assignments 1 – 3	Three assignments, three marks at 10% each = 30% of course marks.
End of course examination	70% of overall course marks
Total	**100% of course materials**

COURSEOVERVIEW

This table brings together the units, the number of weeks you should take to complete them, and the assignments that follow them. These specially designed study materials should be used at your pace, and at a time and place that suit you best. Think of it as reading the lecture instead of listening to a lecturer. The study units tell you when to read your course material. Just as a lecturer might give you an in-class exercise, your study units provide exercises for you to do at appropriate points.

Table 2: Course Organisation

Unit	**Title of Work**	**Week Activity**	**Assessment (End of Unit)**
	Course Guide	**Week**	
1	Basic Statistical Terms and Data Collection Methods	Week 1	Assignment 1
2	Clinical Trial, Epidemiology, Study Designs and Sampling Methods	Week 2	Assignment 2

3	Data presentation: Tabular and Graphical methods	Week 3	Assignment 3
4	Measures of Location, Partition and spread	Week 4	Assignment 4
5	Permutation, Combination and Introduction to Probability	Week 5	Assignment 5
6	Random variables and Probability Distribution	Week 6	Assignment 6
7	Sampling Distribution	Week 7	Assignment 7
8	Estimation	Week 8	Assignment 8
9	Sampling Distributions of the Sample Mean and Proportion	Week 9	Assignment 9
10	Confidence Intervals for population Mean and Proportion	Week 10	Assignment 10
11	Concepts in Testing a Hypothesis	Week 11	Assignment 11
12	Tests for Mean and Proportion	Week 12	Assignment 12
13	One Way ANOVA and Chi Square Test	Week 13	Assignment 13
14	Correlation and Regression Analyses	Week 14	Assignment 14

Each of the study units follows a common format:

i. Introduction to the subject matter of the unit and how a particular unit is integrated with the other units and the course as a whole.

ii. A set of learning objectives. These objectives let you know what you should be able to do by the time you have completed the unit. You should use these objectives to guide your study. When you have finished the unit you must go back and check whether you have achieved the objectives. If you make a habit of doing this, you will significantly improve your chances of passing the course

iii. The main body of the unit guides you through the required reading from other sources. This will usually be either from a reading section or some other courses.

iv. Self-tests are interspersed throughout the units, and answers are given at the end of units. Working through these tests will help you to achieve the objectives of the units and prepare you for the assignments and the examination. You should do each self-test as you encounter it in the study unit. There will also be numerous examples given in the study units; work through these when you come across them too.

If you run into any trouble, telephone your tutor. Remember that your tutor's job is to help you; so when you need help, don't hesitate at all to ask your tutor to provide it.

HOW TO GET THE MOST FROM THIS COURSE

The following is a practical strategy for working through the course.

i. Read this **Course Guide** thoroughly.

ii. Organise a study schedule: Refer to the "course overview" for more details. Note the time you are expected to spend on each unit and how the assignments relate to the units. Important information, e.g. details of your tutorials, and the date of the first day of this material is available. You need to gather all this information in one place, such as your diary or a wall calendar. Whatever method you choose to use, you should decide on and write in your own dates for working on each unit.

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

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

v. Assemble the study materials: Information about what you need for a unit is given on the contents page at the beginning of each unit. You will almost always need both the study unit you are working on and one of the materials for further reading on your desk at the same time.

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

vii. Keep in mind that you will learn a lot by doing all your assignments carefully. They have been designed to help you to meet the objectives of the Course and, therefore will help you pass the exam. Submit all assignments not later than the due date.

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

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

x. When you have submitted an assignment to your tutor for marking do not wait for its return before starting on the next unit. Keep to your schedule. When the assignment is returned, pay particular attention to your tutor's comments, both on the Tutor-Marked Assignment form and also on the written assignment. Consult your tutor as soon as possible if you have any question or problems.

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

FACILITATORS/TUTORS AND TUTORIALS

There are 8 hours of tutorials provided in support of this Course. You will be notified of the dates, times and location of these tutorials, together with the name and phone numbers of your tutor, as soon as you are allocated a tutorial group. Your tutor will mark and comment on your assignment, keep a close watch on your progress and on any difficulties you might encounter and provide assistance to you during the course. You must mail your tutor marked assignments to your tutor well before the due date (at least two working days are required). They will be marked by your tutor and returned to you as soon as possible.

Do not hesitate to contact your tutor by telephone, e-mail for discussion.

Contact your tutor if you:

i. do not understand any part of the study units or the assignment

ii. have difficulty with the self-tests or exercises

iii. have a question or problem with an assignment, with your tutor's Comments on an assignment, or with the grading of an assignment.

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

CONTENTS **PAGE**

MODULE 3 SAMPLING DISTRIBUTIONS AND ESTIMATION

MODULE 4 TEST OF HYPOTHESIS

MODULE ONE INTRODUCTION TO STATISTICS, STUDY DESIGNS AND DATA PRESENTATION

Unit 1: Basic Statistical Terms and Data Collection Methods
Unit 2: Clinical Trial, Epidemiology, Study Design and Sampling Methods

Unit 3: Frequency Distribution Tabular and Graphical Methods

UNIT 1 BASIC STATISTICAL TERMS AND DATA COLLECTION METHODS

CONTENTS

1.0 INTRODUCTION

With the advent of computers and the information age, vast amounts of data are being generated in many fields. The term statistics is used to mean either statistical data or statistical methods. When it means statistical data it refers to numerical descriptions of things. These descriptions may take the form of counts or measurements. Thus statistics of HIV cases, malaria cases, fever cases, number of positives obtained, sex and age distribution of positive cases, etc. When the term 'statistics' is used to mean 'statistical

methods' it refers to a body of methods that are used for collecting, organising, analyzing and interpreting numerical data for understanding a phenomenon or making wise decisions

2.0 OBJECTIVES

After completing this unit, you will be able to:

i. define Statistics Biostatistics and Medical Statistics
ii. define basic terms in Statistics
iii. describe data in different forms and methods of collection of data
iv. describe different methods of data collection

3.0 MAIN CONTENT

3.1 DEFINITION OF STATISTICS, BIOSTATISTICS AND MEDICAL STATISTICS

Statistics is the science of collecting, classifying, presenting, analyzing and interpreting data.

The above definition covers the following:

i. Collection of data: The collection of data is the first step of statistical investigation.
ii. Presentation: After the collection of data, they are presented in systematic form such as table and graphical form.
iii. Analysis: After the collection and presentation of data, the next step is to analyze the data. To analyze the data, we use average, median, regression, etc.
iv. Interpretation: The last step is the interpretation of the results obtained from the analysis and taking appropriate decisions.

The use of statistics allows clinical researchers to draw reasonable and accurate inferences from collected information and to make sound decisions in the presence of

uncertainty. Mastery of statistical concepts can prevent numerous errors and biases in biological or medical research.

Biostatistics is a branch of biological science which deals with the study and methods of collection, presentation, analysis and interpretation of data. Biostatistics is also called biological statistics or biometry. Biostatistics is the application of medicine with statistics, whereas Statistics involves collecting, recording and evaluating data of any type. The former is applied mostly in biological evaluations whereas the latter is utilized to reach at conclusions in each and every field, which involves population. Biostatistics deals with biological problems arising in the health-related sciences and the agricultural sciences. Medical statistics refers to statistics related to biological problems in health-related sciences.

Medical statistics is a sub-discipline of statistics. that involves the science of summarizing, collecting, presenting and interpreting data in medical practice, and using them to estimate the magnitude of associations and test hypotheses.Medical statistics deals with applications of statistics to medicine and the health sciences, including epidemiology, public health, forensic medicine, and clinical research.

However, "biostatistics" more commonly connotes all applications of statistics to biology while Medical statistics is a sub-discipline of statistics. Thus, Medical statistics is the application of statistical knowledge and methods to the field of medicine and medical practice.

3.2 BASIC STATISTICAL TERMS

i. Population: Population is the entire collection, or set, of individuals or objects whose properties are to be analyzed. The population of concern must be carefully defined and is considered fully defined only when its membership list of elements is specified. The set of “all students who took statistics course in year two” is an example of a well-defined

ii. Target population: The population about which we intend to make estimates and inferences

iii. Source population: The population from which cases, controls and samples arise.

iv. Actual population: The population to which our estimates and inferences actually apply

v. Study population: The subjects included in all phases of the study. Population involves not only people but also a collection of animals, manufactured objects, or whatever. There are two types of population, finite and infinite. When the members in a population can be physically counted, the population is said to be finite. It is infinite when the membership is uncountable. The number of students in University of Lagos is a finite population. The set of all registered voters in Nigeria is a very large finite population. On the other hand, the population of all stars in the sky and the population of all sands at the seashore all over the world are infinite.

vi. Census: Survey of the entire population. It refers to a nation-wide counting of population or Complete enumerations.

vii. Survey: A descriptive study done generally on scientifically selected subjects. Another descriptive study methodology is case series.

viii. Sample: A sample is a subset of a population. A subset consists of the individuals, objects, or measurements selected by the sample collector from the population. For example, a set of males in the department of Physiotherapy is a subset of the number of students in the department. A set of Toyota cars parked at the faculty car park is a subset of cars parked at the faculty car park.

ix. Statistic: A statistic is a quantity whose numerical values can be obtained from sample data. A statistic is a value that describes a sample. Sample statistics, for example include: mean, median, mode etc.

x. Parameter: A parameter is a quantifiable characteristic of a population, the value of which is needed to completely specify the distribution of the reference population. Suppose we want to determine the value of the population mean (μ).

xi. Variable: A characteristic of interest about each individual element of a population or sample. A student's name, matriculation number, year and department are all variables.

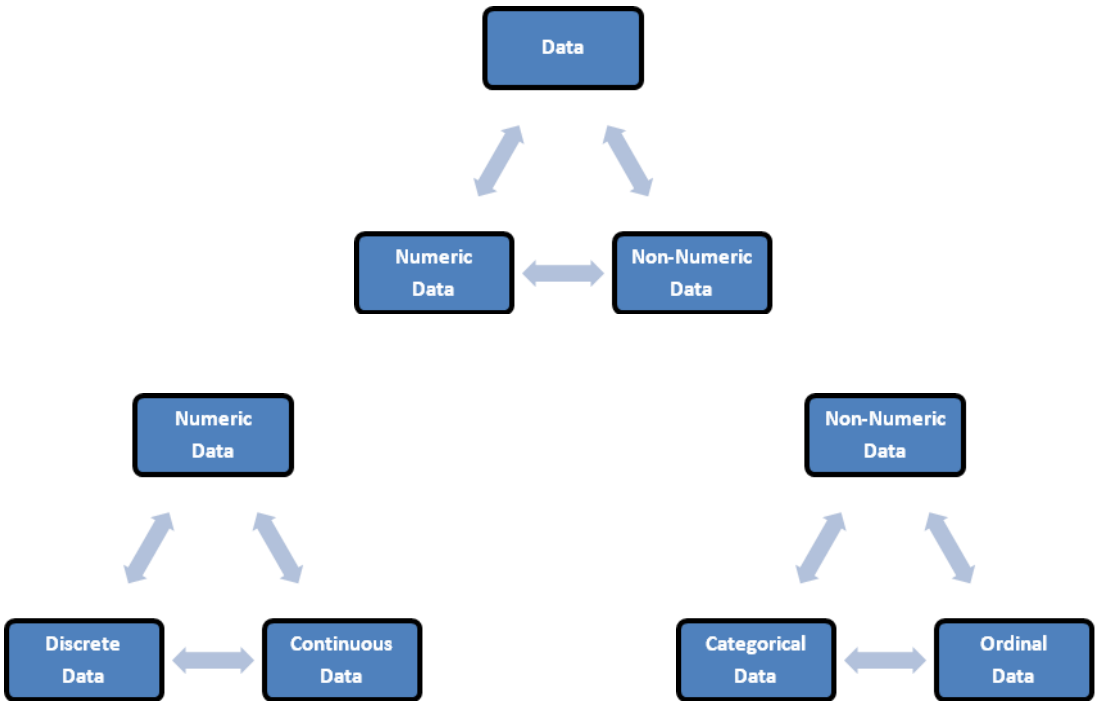
xii. Experiment: A planned activity whose results yield set of data is known as experiment.

3.3 DATA AND DATA COLLECTION METHODS

These are raw facts or unprocessed information. There are basically two types of data:

i. numeric or quantitative information and

ii. non-numeric or qualitative information.



There are two classifications of numeric data:

i. Discrete numeric data: These are countable whole numbers or integers, e.g. the number of Patients, nurses or doctors in a hospital
ii. Continuous numeric data: These are numbers within an interval or uncountable range of numbers, e.g. A measure of a quantity will usually be continuous, i.e. weight, mass, litres etc.

Non-numeric data are values that cannot be quantified. For example, matriculation number, age group, blood group, DNA, RNA, sex, tribe, country, etc. Data in this form are either categorical or ordinal. Examples of ordinal non-numeric data are students' height, age group, while DNA, RNA, sex, country, tribe are examples of categorical non-numeric data.

Data are also expressed using the following scales of measurement: nominal, ordinal, interval, and ratio. The scale of measurement determines the amount of information contained in the data and indicates the data summarization and statistical analyses that are most appropriate.

Nominal scale: when the data are labels or names used to identify an attribute of the element. There is no implied order to the categories of nominal data. In these types of data, individuals are simply placed in the proper category or group, and the number in each category is counted. Each item must fit into exactly one category.

Ordinal Data: have order among the response classifications (categories). The spaces or intervals between the categories are not necessarily equal. Example: strongly agree, agree, no opinion, disagree, strongly disagree. Here, we can see that the data are ordered.

Interval Data: In interval data the intervals between values are the same. For example, in the Fahrenheit temperature scale, the difference between 70 degrees and 71 degrees is the same as the difference between 32 and 33 degrees. But the scale is not a ratio scale. 40 degrees Fahrenheit is not twice as much as 20 degrees Fahrenheit.

Ratio Data: The data values in ratio data do have meaningful ratios, for example, age is a ratio data, and someone who is 40 is twice as old as someone who is 20.

Data collected for investigation can either be primary data or secondary data. Data collected directly from the source or respondents are known as primary data. These are data, which are collected by the investigator for the purpose of a specific inquiry or study. Such data are original in character and are mostly generated by surveys conducted by individuals or research institutions. When an investigator uses data, which have already been collected by others or those from established data bank are known as secondary data. Such data are primary data for the agency that collected them, and become secondary for someone else who uses these data for his own purposes. Secondary data are less expensive to collect both in money and time.

Data collection techniques allow us to collect data about our objects of study (people, objects, and phenomena) and about the setting in which they occur. In the collection of data, we have to be systematic otherwise it will be difficult to answer our research questions in a conclusive way.

How to Collect Primary Data

i. Observation: Observation is a technique that involves systematically selecting, watching and recoding behaviors of people or other phenomena and aspects of the setting in which they occur, for the purpose of getting (gaining) specified information.
ii. Face-to-face and Telephone interviews: A good interviewer can stimulate and maintain the respondent's interest, and can create a rapport (understanding, concord) and atmosphere conducive to the answering of questions.
iii. Experiments
iv. Self-administered Questionnaire

Sources of secondary data include:

i. Official publications of Central Statistical Authority
ii. Publication of Ministry of Health and Other Ministries
iii. Internet, Websites, News Papers and Journals

iv. International Publications like Publications by WHO, World Bank, UNICEF
v. Records of hospitals or any Health Institutions.

Problems of Data Collection

The main problems that may be faced when collecting data are in the selection of appropriate collection methods and in the training of the staff involved. Other problems include Language barriers, Lack of adequate time, Expense, inadequately trained and experienced staff, Invasion of privacy, Suspicion, Bias, Cultural norms.

Design of Questionnaire

Before examining the steps in designing a questionnaire, we need to review the types of questions used in questionnaires. Depending on how questions are asked and recorded we can distinguish between two major possibilities: Open-ended questions and Closed questions.

Open-ended questions permit free responses that should be recorded in the respondent's own words. The respondent is not given any possible answers to choose from.

Closed questions offer a list of possible options or answers from which the respondents must choose.

Census

Ordinarily, the term "census" refers to a nation-wide counting of population. Complete enumerations or censuses are taken by obtaining information concerning every inhabitant of an area. Information to be collected include: Sex, age, marital status, educational characteristics, economic characteristics, place of birth, language, fertility mortality, citizenship (nationality), living conditions (e.g. house-ownership, type of housing and the like), religion, etc. In short, the main characteristics of a census could be summarized as follows:

i. Separate enumeration and recording of the characteristics of each individual
ii. It should refer to people inhabiting a well-defined territory
iii. The population should be enumerated with respect to a well-definedpoint in time

iv. It should be taken at regular intervals (usually every ten years). In most countries the personal data collected in a census are not used for other than statistical purposes.

v. The compilation and publication of data by geographic areas and by basic demographic variables is an integral part of a census.

4.0 CONCLUSION

The definition of statistics, biostatistics or medical statistics and description of statistical data and methods are accompanied by examples illustrating uses. Basic foundation for understanding the concepts of statistics in field of medicine established. The different forms of data and methods of data collection enumerated.

5.0 SUMMARY

This module has covered definitions of Statistics, Biostatistics and Medical Statistics and the relationship between them. It has also covered basic definitions of terms relevant to understanding of concepts in the contents. Data have been extensively broken down into its constitute units and elaborated. The different forms of data and methods of collection of data

6.0 TUTOR-MARKED ASSIGNMENTS

1.(a) Differentiate between Statistics and Biostatistics

(b) Define and explain the following terms:

(i) Population (ii) Sample (iii) Statistic (iv) Statistics (v) Variable

(vi) Data (vii) Experiment

2.(a). Select twenty students currently enrolled in your department and collect data for these three variables; number of courses enrolled in, total cost of textbooks and method of payment used for textbooks

(b) What is the population? Is the population finite or infinite? What is the sample?

(c.) Classify the responses for each of the three variables as non-numeric data, discrete data, or continuous data.

3.Classify each of the following as examples of (i) Non-numeric (ii) Discrete (iii) Continuous variables:

i. The hair colour of people in a concert show.
ii. The number of hours required to heal a patient of a disease.
iii. The length of time required answering a telephone call at a certain business center.
iv. The number of pages per job coming off a computer printer.
v. The kind of trees used as Christmas tree.
vi. The number of voters in a community.
vii. Whether a statement is true or false.
viii. The number of books in a library.

REFERENCES AND FURTHER READINGS:

Department of Mathematics (2015). A First Course in Statistics. Department of Mathematics, University of Lagos, Akoka-Yaba, Lagos, Nigeria.

Degu G. and Tessema F. (2007). Biostatistics. In collaboration with the Ethiopia Public Health Training Initiative, The Carter Center, the Ethiopia Ministry of Health, and the Ethiopia Ministry of Education.

Indrayan A. (2017). Statistical medicine: An emerging medical specialty. J Postgrad Med. Available from: http://www.jpgmonline.com/text.asp?2017/63/4/252/216438. Volume 63, (4) 252 – 256.

National Library of Medicine. Epidemiology Studies. National Institutes of Health, 9000 Rockville Pike, Bethesda, Maryland 20892U.S. Department of Health and Human Services. https://toxtutor.nlm.nih.gov/05-003.html.

Petrie A. and Sabin C.(2005). Medical Statistics at a Glance. Published by Blackwell Publishing Ltd, USA.

Song, J. W., & Chung, K. C. (2010). Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, *126*(6), 2234–2242. https://doi.org/10.1097/PRS.0b013e3181f44abc

Study Design 101 by Himmelfarb Health Sciences Library. 2011-2019, The Himmelfarb Health Sciences Library.

UNIT 2: CLINICAL TRIAL, EPIDEMIOLOGY, STUDY DESIGNS AND SAMPLING METHODS

CONTENTS

1.0 INTRODUCTION

Statistical analysis is an essential technique that enables a medical research practitioner to draw meaningful inference from their data analysis. Improper application of study design

and data analysis may render insufficient and improper results and conclusion. There are several types of epidemiology/clinical trial design. These can be classified according to the method used to allocate participants into treatment or control groups, according to the awareness of either participants or researchers or both, of which group participants are allocated into single or double-blind studies according to the magnitude of difference between treatment and control groups that is expected.

2.0 OBJECTIVES

After completing this module, you will be able to:

i. Define and differentiate between Clinical Trial and Epidemiology
ii. Define and differentiate between Experimental and Observational Studies
iii. Understand the design, conduct and interpretation of clinical and epidemiological studies
iv. Understand Cross-sectional studies, Cohort studies, Case-control studies, Longitudinal studies
v. Define Sampling and describe common sampling methods

3.0 MAIN CONTENT

3.1 CLINICAL TRIAL AND EPIDEMIOLOGY

A Clinical trial is a medical experiment on human subjects, particularly in a clinic setup, such as to find efficacy and safety of a new therapeutic or diagnostic regimen. A clinical trial actually is an experiment testing medical treatments on human subjects. The clinical investigator controls factors that contribute to variability and bias such as the selection of subjects, application of the treatment, evaluation of outcome, and methods of analysis. The distinction of a clinical trial from other types of medical studies is the experimental nature of the trial and its occurrence in humans. We defined a clinical trial as a prospective study that compares the effects and value of intervention (s) against a control in human beings. Note that a clinical trial is prospective, rather than retrospective.

A clinical trial must employ one or more intervention techniques. These may be single or combinations of diagnostic, preventive, or therapeutic drugs, biologics, devices, regimens, procedures, or educational approaches. Intervention techniques should be applied to participants in a standard fashion in an effort to change some outcome. Follow-up of people over a period of time without active intervention may measure the natural history of a disease process, but it does not constitute a clinical trial. Without active intervention the study is observational because no experiment is being performed.

The term “clinical trial” is preferred over “clinical experiment” because the latter may connote disrespect for the value of human life.

Epidemiology is the study of factors that affect distribution and determinants of disease or a health condition in a human population. Epidemiology is the study of diseases in populations of humans or other animals, specifically how, when and where they occur. Epidemiologists attempt to determine what factors are associated with diseases (risk factors), and what factors may protect people or animals against disease (protective factors).

Epidemiology studies are conducted using human populations to evaluate whether there is a correlation or causal relationship between exposure to a substance and adverse health effects.

These studies differ from clinical investigations in that individuals have already been administered the drug during medical treatment or have been exposed to it in the workplace or environment.

Epidemiological studies measure the risk of illness or death in an exposed population compared to that risk in an identical, unexposed population (for example, a population the same age, sex, race and social status as the exposed population).

Standard and quantitative measures are used to determine if epidemiological data are meaningful. This can be achieved using some common statistical measures:

i. **Odds Ratio (O/R)**: The ratio of risk of disease in a case-control study for an exposed group to an unexposed group. An odds ratio equal to 2 (O/R = 2) means that the exposed group has twice the risk as the non-exposed group.

ii, **Standard Mortality Ratio (SMR):** The relative risk of death based on a comparison of an exposed group to non-exposed group. A standard mortality ratio equal to 150 (SMR = 150) indicates that there is a 50% greater risk.

iii. **Relative Risk (RR)**: The ratio expressing the occurrence of disease in an exposed population to that of an unexposed population. A relative risk of 175 (RR = 175) indicates a 75% increase in risk.

Epidemiologists attempt to control errors that can occur in the collection of data, which are known as bias errors. The three main types of bias errors are:

i. Selection bias: This occurs when the study group is not representative of the population from which it came.

ii. Information bias: This occurs when study subjects are misclassified as to disease or exposure status. Recall bias occurs when individuals are asked to remember exposures or conditions that existed years before.

ii. Confounding factors: which occur when the study and control populations differ with respect to factors which might influence the occurrence of the disease. For example, smoking might be a confounding factor and should be considered when designing studies.

Often, however, epidemiology provides sufficient evidence to take appropriate control and prevention measures. Epidemiologic studies fall into two categories: experimental and observational.

3.2 EXPERIMENTAL, OBSERVATIONAL AND LONGITUDINAL STUDIES

3.2.1 EXPERIMENTAL AND OBSERVATIONAL STUDIES

An experimental study is a study where the researcher has control over most of the variables. Once the research problem has been formed, the researcher organizes a study

that will allow him to find answers to the research problem. In this case, the researcher conducts the study in a specific setting such as a laboratory where he can control the variables. This, however, does not entail that all variables can be controlled. On the contrary, some variables can be beyond the control of the researcher. Experimental studies are mainly conducted in the natural sciences. This does not denote that experimental studies cannot be conducted in the social sciences.

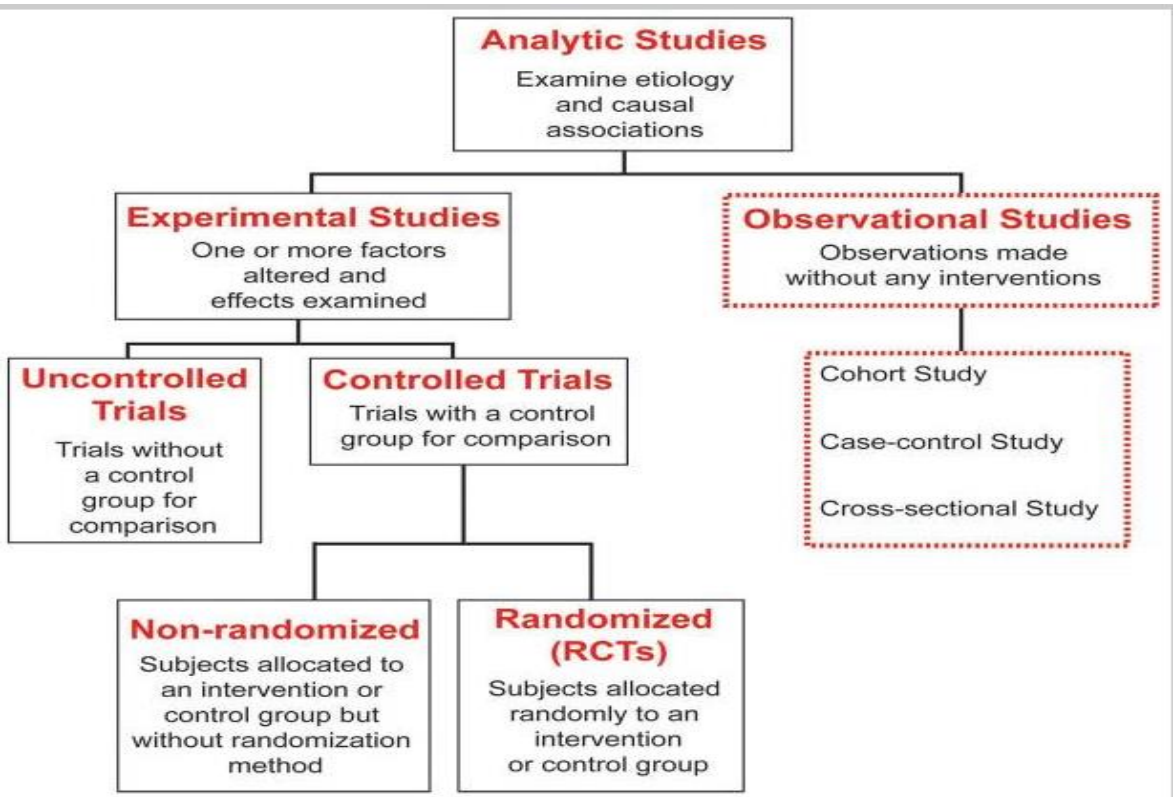
An observational study is a study where the researcher merely observes the subject without controlling any variables. These types of studies are mainly used in the social sciences. In disciplines such as sociology, anthropology, etc. observational studies are used to comprehend human behavior. Observational studies can also be conducted in the natural sciences as well in order to comprehend behavioral patterns.

Basic Differences between Experimental and Observational Studies

Experimental Studies	Observational Studies
i) An experimental study is a study where the researcher has control over most of the variables. ii) The researcher has control over the variables. He can manipulate variables in order to make changes in the environment. iii) Experimental studies are mostly conducted in the natural sciences. vi) The laboratory setting is mostly suitable since variables can be easily controlled.	i) An observational study is a study where the researcher merely observes the subject without controlling any variables ii) The researcher does not control the research environment, he merely observes. iii) Observational studies are mostly conducted in the social sciences. iv) The natural setting is used, where the research subjects can act naturally without being controlled.

3.2.2 STUDY DESIGNS

Study design is the planning or designing of a research study. Successful studies need to address two important dimensions: reliability and validity. A reliable study should be replicable, provide similar results if the same study parameters are applied. Validity is concerned with the ability of the study to correctly answer the question it asks.



Source: Song, J. W., & Chung, K. C. (2010)

Three major types of epidemiologic studies are cohort, case-control, and cross-sectional studies. A cohort, or longitudinal study follows a defined group over time.

Cohort Study

A study design where one or more samples (called cohorts) are followed prospectively and subsequent status evaluations with respect to a disease or outcome are conducted to determine which initial participants' exposure characteristics (risk factors) are associated with it. As the study is conducted, outcome from participants in each cohort is measured

and relationships with specific characteristics determined. There are two types of cohort studies:

Prospective, in which cohorts are identified based on current exposures and followed into the future.

Retrospective, in which cohorts are identified based on past exposure conditions and study "follow-up" proceeds forward in time; data come from past records.

Advantages

i. Subjects in cohorts can be matched, which limits the influence of confounding variables
ii. Standardization of criteria/outcome is possible
iii. Easier and cheaper than a randomized controlled trial (RCT) Disadvantages
iv. Cohorts can be difficult to identify due to confounding variables
v. No randomization, which means that imbalances in patient characteristics could exist
vi. Blinding/masking is difficult
vii. Outcome of interest could take time to occur

Case Control Study

A study that compares patients who have a disease or outcome of interest (cases) with patients who do not have the disease or outcome (controls), and looks back retrospectively to compare how frequently the exposure to a risk factor is present in each group to determine the relationship between the risk factor and the disease.

Case control studies are observational because no intervention is attempted and no attempt is made to alter the course of the disease. The goal is to retrospectively determine the exposure to the risk factor of interest from each of the two groups of individuals: cases and controls. These studies are designed to estimate odds. Case control studies are also known as "retrospective studies" and "case-referent studies."

Advantages

i. Good for studying rare conditions or diseases

ii. Less time needed to conduct the study because the condition or disease has already occurred
iii. Let's you simultaneously look at multiple risk factors
iv. Useful as initial studies to establish an association
v. Can answer questions that could not be answered through other study designs

Disadvantages

i. Retrospective studies have more problems with data quality because they rely on memory and people with a condition will be more motivated to recall risk factors (also called recall bias).
ii. Not good for evaluating diagnostic tests because itâ€™s already clear that the cases have the condition and the controls do not
iii. It can be difficult to find a suitable control group
iv. Design pitfalls to look out for
v. Care should be taken to avoid confounding, which arises when an exposure and an outcome are both strongly associated with a third variable. Controls should be subjects who might have been cases in the study but are selected independent of the exposure. Cases and controls should also not be "over-matched."

Cross-Sectional Studies

A cross-sectional study is a type of observational research that analyzes data of variables collected at one given point in time across a sample population or a pre-defined subset. This study type is also known as cross-sectional analysis, transverse study, or prevalence study. Although cross-sectional research does not involve conducting experiments, it is often used to understand outcomes in the physical and social sciences, as well as many business industries.

Cross-sectional research study is either descriptive or analytical. A descriptive cross-sectional study assesses how frequently, widely, or severely the variable of interest occurs throughout a specific demographic. Analytical cross-sectional research investigates the association between two related or unrelated parameters. This

methodology isn't entirely foolproof, though, because the presence of outside variables and outcomes are simultaneous, and their studies are, too. In a real-life cross-sectional study, researchers usually use both descriptive and analytical research methods.

Examples

i. Scientists in healthcare may use cross-sectional research to understand how children ages 2-12 across the Nigerian States are prone to vitamin A deficiency.
ii. A medical study looking at the prevalence of cancer amongst a defined population. The researcher can evaluate people of different ages, ethnicities, geographical locations, and social backgrounds. If a significant number of men from a particular age group are more prone to have the disease, the researcher can conduct further studies to understand the reasons behind it (longitudinal study), which would study the same participants over time.

Advantages of cross-sectional studies

i. Relatively quick to conduct.
ii. All variables are collected at one time.
iii. Multiple outcomes can be researched at once.
iv. Prevalence for all factors can be measured.
v. Suitable for descriptive analysis.
vi. Can be used as a springboard for further research.

Disadvantages of cross-sectional studies

i. Cannot be used to get timeline-based research.
ii. Tough to find people that fall under the same variables.
iii. Associations can be difficult to interpret.
iv. Is open to bias.
v. Does not help to determine cause.

Randomized Controlled Trial

This is a study design that randomly assigns participants into an experimental group or a control group. As the study is conducted, the only expected difference between the control and experimental groups in a randomized controlled trial (RCT) is the outcome variable being studied. The variables being studied should be the only variables between the experimental group and the control group.

Advantages

i. Good randomization will "wash out" any population bias
ii. Easier to blind/mask than observational studies
iii. Results can be analyzed with well-known statistical tools
iv. Populations of participating individuals are clearly identified

Disadvantages

i. Expensive in terms of time and money
ii. Volunteer biases: the population that participates may not be representative of the whole
iii. Loss to follow-up attributed to treatment
iv. Design pitfalls to look out for
v. An RCT should be a study of one population only.

3.2.3 LONGITUDINAL STUDIES

Longitudinal studies are versatile, repeatable, and often use surveys to collect data that is either qualitative or quantitative. Additionally, in a longitudinal study, a survey creator does not interfere with survey participants. Instead, the survey creator distributes questionnaires over time to observe changes in participants, their behaviours, or their attitudes. The purpose of using the same individuals or samples in the longitudinal study is to observe any measurable change over time. This ensures that you can account for the same variables of interest in the duration of your study.

Example

Consider a study conducted to understand the similarities or differences between identical twins who are brought up together versus identical twins who were not. The study observes several variables, but the constant is that the participants all have an identical twin. Researchers, in this case, would want to observe these participants from childhood to adulthood to understand how growing up in different environments influences traits, habits, and personality. Over many years, researchers can see both sets of twins as they experience life without intervention. Because the participants share the same genes, it is assumed that any differences are due to environmental factors, but only attentive study can bring them to a conclusion.

There are three major types of longitudinal studies for future research:

i. Panel study: A panel study involves a sample of people from a bigger population and is conducted at specified intervals for a longer period. One of the most important features of the panel study is that data is repeatedly collected from the same sample at different points in time. Most panel studies are designed for quantitative analysis, though they are also used for collecting qualitative data and analysis.

ii. Cohort Study: A cohort study samples a cohort (a group of people who typically experience a common event at a given point in time). Medical researchers tend to conduct cohort studies. Some might consider clinical trials similar to cohort studies. However, in cohort studies, researchers merely observe participants without intervention, unlike clinical trials in which participants undergo tests.

iii. Retrospective study: A retrospective study makes use of already existing data, collected during previously conducted research with similar methodology and variables. While conducting a retrospective study, the researcher uses an administrative database, pre-existing medical records, or one-to-one interviews.

Longitudinal study is useful in science and medicine as well as many other fields. There are many reasons why a researcher might want to conduct a longitudinal study. One of

the important reasons is, longitudinal studies give unique insights that many other types of research fail.

Advantages of Longitudinal Study

i. Longitudinal study can identify and relate to events. You can reveal chronology between events like long-term and short-term changes in variables, making this ideal for medical studies.
ii. Similarly, because a longitudinal study is carried out over a long period, it helps identify and establish a particular sequence of events.
iii. Longitudinal study provides meaningful insights that might not be possible with other forms of research like cross-sectional and similar studies.
iv. Longitudinal study allows researchers to trace development over a timeline instead of drawing conclusions based on a "snapshot" of data.

Disadvantages of Longitudinal Study

i. Longitudinal studies are not cost-effective. Because they can run long, the costs can add up.
ii. An extended period may lead to longitudinal survey respondents dropping out during the study.
iii. Participants may start to act unnaturally because they know they are under observation, which spoils the research.
iv. Continuity over the years may be challenging to maintain. For example, if the lead researcher of the study retires, the person replacing them may or may not have the same report.

3.3 PREVALENCE, INCIDENCE AND SCREENING TEST

Prevalence

Prevalence is a measure of disease that allows us to determine a person's likelihood of having a disease. Therefore, the number of prevalent cases is the total number of cases of disease existing in a population. A prevalence rate is the total number of cases of a disease existing in a population divided by the total population. So, if a measurement of

cancer is taken in a population of 40,000 people and 1,200 were recently diagnosed with cancer and 3,500 are living with cancer, then the prevalence of cancer is 0.118. (or 11,750 per 100,000 persons)

Prevalence gives a figure for a factor (disease, injury, health status etc) at a single point in time (point prevalence) or time period (period prevalence). Period prevalence provides the better measure of the factor since it includes all cases between two dates, whereas point prevalence only counts cases on a particular date. It is a measure of disease that allows us to determine a person's likelihood of having a disease. It is most meaningfully reported as the number of cases as a fraction of the total population at risk and can be further categorised according to different subsets of the population.

An example of prevalence: A recent Scottish study showed that the prevalence of obesity in a group of children aged from 3 to 4 years was 12.8% at the time.

Incidence is often confused with prevalence. The easy way to remember the difference is that prevalence is the proportion of cases in the population at a given time rather than rate of occurrence of new cases. Thus, incidence conveys information about the risk of contracting the disease, whereas prevalence indicates how widespread the disease is.

Incidence

Incidence is a measure of disease that allows us to determine a person's probability of being diagnosed with a disease during a given period of time. Therefore, incidence is the number of newly diagnosed cases of a disease. An incidence rate is the number of new cases of a disease divided by the number of persons at risk for the disease. If, over the course of one year, five women are diagnosed with breast cancer, out of a total female study population of 200 (who do not have breast cancer at the beginning of the study period), then we would say the incidence of breast cancer in this population was 0.025. (or 2,500 per 100,000 women-years of study)

Mortality

Mortality is another term for death. A mortality rate is the number of deaths due to a disease divided by the total population. If there are 25 lung cancer deaths in one year in a population of 30,000, then the mortality rate for that population is 83 per 100,000.

Screening Test

A simple test performed on a large number of people to identify those who have or are likely to develop a specified disease.

A preliminary or abridged test intended to eliminate the less probable members of an experimental series.

3.4 SAMPLING AND COMMON SAMPLING METHODS

The process by which samples are obtained from a population is called sampling.

There are two main types of sampling: probability and non-probability sampling. The difference between the two types is whether or not the sampling selection involves randomization. Randomization occurs when all members of the sampling frame have an equal opportunity of being selected for the study. Following is a discussion of probability and non-probability sampling and the different types of each.

i. Probability Sampling – Uses randomization and takes steps to ensure all members of a population have a chance of being selected. The different types of probability sampling include simple random sampling, stratified sampling, cluster sampling, multistage sampling, and systematic random sampling
ii. Non-probability Sampling – Does not rely on the use of randomization techniques to select members. This is typically done in studies where randomization is not possible in order to obtain a representative sample. Bias is more of a concern with this type of sampling. The different types of non-probability sampling include convenience or accidental sampling, purposive sampling, expert sampling, quota sampling and snowball sampling.

3.4.1 Probability Sampling Methods

Simple Random Sampling

A sample selected in such a way that every element in the population has an equal probability of being chosen is known as a random sample. Random samples are obtained by either sampling with replacement from a finite population or by sampling without replacement from an infinite population. To select a simple random sample, first assign a number to each element in the sampling frame. This is usually done sequentially using the same number of digits for each element. Then pick as many numbers from the table of random numbers with that number of digits as are needed for the sample size desired.

Systematic Sampling

One of the easiest-to-use methods for approximating a random sample is the systematic sampling method. This is a sample in which every k^{th} item in the sampling frame is selected after a random start among the first K elements.

For example, if we desire a 10% systematic sampling, we would determine the position of the first element by using the random number table to randomly select a number between 1 and 10. Suppose the random number table gives 6, then the first element to be picked would come from the 6^{th} position on the sampling frame. Then the second will be from the 16^{th} position, the third from the 26^{th} position and so on.

Stratified Sampling

When sampling very large populations, sometimes it is possible to divide the population into sub-populations on the basis of some characteristics. These sub-populations are called strata. A stratified sampling is obtained by stratifying the sampling frame and then selecting a fixed number of items from each of the strata by means of random sampling.

When a stratified sample is to be drawn, the population is subdivided into the various strata and then a sub-sample is drawn from each stratum. These sub-samples may be drawn from each stratum randomly or systematically. Then the sub-samples are summarized separately and this information is combined to draw conclusions about the whole population. Examples of this include:

i. Students in a faculty stratified by their department
ii. Income groups – upper, middle and low-income groups.

Cluster Sampling

A cluster sampling is obtained by stratifying the sampling frame and then selecting all of the items from some, but not all, of the clusters. The cluster sample is obtained by using either random numbers or a systematic method to identify the strata (clusters) to be sampled and then using all the items from within these clusters. Then the sub-samples are summarized separately and this information is combined.

For example, to study the living condition in the six geopolitical regions in Nigeria, three geopolitical regions are randomly selected. Then all units in the cluster sample are studied. This study of all the units makes cluster sampling sometimes unattractive.

Two-Stage Sampling

A two-stage sampling is obtained at two stages. The first stage involves the division of the population into what is known as primary sampling units (PSU's). This first sampling involves a group of individuals in the population put on the basis of their closeness or proximity. A frame of secondary units (SU's) is selected from the PSU's corresponding to individuals in the population, through random sampling or systematic sampling. For example, in a study of Universities in Lagos State, the primary sampling units can be the University and the secondary units the departments in the University. This can be extended to three of more stage sampling. For example, the third stage would involve the students in the university.

3.4.2 Non Probability Sampling Methods or Non Random Sampling Methods

Quota Sampling

Samples that are selected on the basis of being "typical" and with no definite probability law associated with the selection procedure are known as quota sampling. When a quota sample is drawn, the person selecting the sample chooses items that he or she thinks are representative of the population. The validity of the results from a quota sampling reflects the soundness of the collector's judgment. For example, when a study of people's idea of a government policy is carried out with the restriction to conduct interview on 100 people each from three major languages; Yoruba, Ibo and Hausa.

Others are:

i. Judgment or Purposive sampling

ii. Convenience sampling

4.0 CONCLUSION

This module has covered various study designs known to clinical trial and epidemiology. Fundamental terms and concepts are also covered.

5.0 SUMMARY

In this unit, medical terms associated with statistics were defined and expatiated. The basic terms include: Clinical, Epidemiological, Cross-sectional studies, Cohort studies, Case-control studies, Longitudinal studies etc. Sampling and common sampling methods were discussed.

6.0 TUTOR-MARKED ASSIGNMENT

1. Definition the following terms:

i. Incidence

ii. Prevalence

iii. Screen Test

iv. Clinical Trial

v. Epidemiological Studies

2. Differentiate between:

i. Clinical Trial and Epidemiological Studies

ii. Experimental studies and Observational studies

iii. **Panel study and Retrospective study**

iv. Case-control studies and Cross-sectional studies.

v. Probability sampling and Non-Probability sampling

References and Further Readings:

Department of Mathematics (2015). A First Course in Statistics. Department of Mathematics, University of Lagos, Akoka-Yaba, Lagos, Nigeria.

Degu G. and Tessema F. (2007). Biostatistics. In collaboration with the Ethiopia Public Health Training Initiative, The Carter Center, the Ethiopia Ministry of Health, and the Ethiopia Ministry of Education.

Indrayan A. (2017). Statistical Medicine: An emerging medical specialty. J Postgrad Med. Available from: http://www.jpgmonline.com/text.asp?2017/63/4/252/216438. Volume 63, (4) 252 – 256.

National Library of Medicine. Epidemiology Studies. National Institutes of Health, 9000 Rockville Pike, Bethesda, Maryland 20892U.S. Department of Health and Human Services. https://toxtutor.nlm.nih.gov/05-003.html.

Petrie A. and Sabin C.(2005). Medical Statistics at a Glance. Published by Blackwell Publishing Ltd, USA.

Song, J. W., & Chung, K. C. (2010). Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, *126*(6), 2234–2242. https://doi.org/10.1097/PRS.0b013e3181f44abc

Study Design 101 by Himmelfarb Health Sciences Library. 2011-2019, The Himmelfarb Health Sciences Library.

Suresh, K., Thomas, S. V., & Suresh, G. (2011). Design, data analysis and sampling techniques for clinical research. *Annals of Indian Academy of Neurology*, *14*(4), 287–290. https://doi.org/10.4103/0972-2327.91951

UNIT 3: DATA PRESENTATION: TABULAR AND GRAPHICAL METHODS

CONTENTS

1.0 INTRODUCTION

One of the first things that you may wish to do when you have entered your data onto a computer is to summarize them in some way so that you can get an insight into the vital information about the dataset. This can be done by producing diagrams, tables or summary statistics. Diagrams are often powerful tools for conveying information about the data, for providing simple summary pictures.

2.0 OBJECTIVES

After completing this module, you will be able to:

i. Identify the different methods of data presentation

ii. Be able to construct Frequency Distribution Tables for both grouped and ungrouped data
iii. Identify the different Graphical (pictorial) representation of the data

For data to be more easily appreciated and to draw quick comparisons, it is often useful to arrange the data in the form of a table or graphical forms. This can be accomplished with the aid of a Frequency distribution table and graphical exploratory data analysis methods.

3.0 MAIN CONTENT

3.1 FREQUENCY DISTRIBUTION TABULAR METHODS

The presentation of data in a meaningful way is done by preparing a frequency distribution. A table listing all classes and their frequencies is called frequency distribution.

Let us demonstrate the concept of a frequency distribution by using the following set.

1	5	3	4	1	3	2	5
2	4	1	3	2	0	1	2
1	2	0	2	1	4	5	3

Let x represent these data values, we can use a frequency distribution to represent this set of data by listing the x values with their frequencies in Table 3.1.

Table 3.1 Frequency distribution

X	0	1	2	3	4	5
F	2	6	6	4	3	3

Group Frequency Distribution

In the case where many different entries for x and several low frequencies, it often makes sense to combine the data in groups or classes. This kind of frequency distribution is called grouped frequency distribution.

Let us demonstrate this with this example:

55	60	61	35	41	43	50	78	72	83
45	70	76	31	49	65	79	83	41	86
53	62	52	47	38	57	64	78	47	54
43	73	85	48	66	48	85	86	82	48
56	84	37	57	57	45	95	45	73	39

The following guidelines and terminology will be used to group continuous-type data into classes of equal length. These guidelines can also be used for sets of discrete data that have a large range.

4.1 Determine the largest (maximum) and smallest (minimum) observations. The range is the difference, R = maximum – minimum

4.2 A frequency distribution should have a minimum of 5 classes and a maximum of 20. For small data sets, use between 5 and 10 classes. For large data sets, use up to 20 classes.

4.3 Each data entry must fall into one and only one class.

4.4 There should be no gaps. Moreover, if there are no entries for a particular class, that class must still be included with a frequency of 0.

4.5 The first interval should begin about as much below the smallest value as the last interval ends above the largest.

The intervals are called class intervals and the boundaries are called class boundaries.

The class limits are the smallest and largest possible observed values in a class.

The class mark is the midpoint of a class.

We set up the following classes for the above data 30 – 39, 40 – 49, 50 – 59, etc. We now create a summary table below:

Table 3.2

Class	Class limits	Frequency	Class boundaries	Class center
1	30 – 39	5	29.5 – 39.5	34.5
2	40 – 49	13	39.5 – 49.5	44.5
3	50 – 59	9	49.5 – 59.5	54.5

4	60 – 69	6	59.5 – 69.5	64.5
5	70 – 79	8	69.5 – 79.5	74.5
6	80 – 89	8	79.5 – 89.5	84.5
7	90 – 99	1	89.5 – 99.5	94.5

Tables like this show us how the data are spread out or distributed; we call this a frequency distribution table or simply a frequency distribution.

Cumulative and Relative Frequencies

When frequencies of two or more classes are added up, such total frequencies are called Cumulative Frequencies. The cumulative frequency of a class is the total of all class frequencies up to and including the present class. On the other hand, relative frequencies express the frequency of each value or class as a percentage to the total frequency.

The relative frequency for a class is the number of entries in the class divided by the total number of entries. It is calculated as

$$\text{Relative frequency} = \frac{\text{frequency in the class}}{\text{Total number of observations}}$$

$$\text{Percentage Relative frequency} = (\text{relative frequency} \times 100)\%$$

For example, the relative frequency of class 50 – 59 is

$$\frac{9}{50} \times 100\% = 18\%$$

The cumulative frequency distribution of the example given above is as follows:

Table 3.3: Cumulative Frequency Table

Class	Class limits	Freq.	Cum Freq.	% Relative Cum. Freq.
1	30 – 39	5	5	10%
2	40 – 49	13	18	36%
3	50 – 59	9	27	54%
4	60 – 69	6	33	66%
5	70 – 79	8	41	82%
6	80 – 89	8	49	98%
7	90 – 99	1	50	100%

Relative Cumulative Frequency is also called Percentage Cumulative Frequency. For example, the Relative Cumulative Frequency for class 60 – 69 is

$$(33/66) \quad x \quad 100\% = 50\%$$

3.2 GRAPHICAL METHODS

Graphical (pictorial) representation of the data reveals patterns of behaviour of the variable being studied. There are several graphic (pictorial) ways to describe data. The type of data and the idea to be presented determines the method used.

Data can be presented graphically in many ways as, line graph, dot plot display, bar chart, pie chart, histogram, cumulative frequency curve (Ogive) and stem-and-leaf display.

Dot Plot Display

Dot plots display the data of a sample by representing each piece of data with a dot positioned along a scale. This scale can be either horizontal or vertical. The frequency of the values is represented along the other scale. They are usually used to represent the frequency distribution of a discrete variable. The dot plot display is a convenient technique to use as you first begin to analyze the data. It results in a picture of the data as well as sorts the data into numerical order.

Example

A random sample of 20 children took their weights in kilogram in a hospital are presented below:

23	22	26	28	22	29	30	25	26	27
21	23	27	26	25	29	30	26	25	28

Construct a dot plot of these data.

Solution

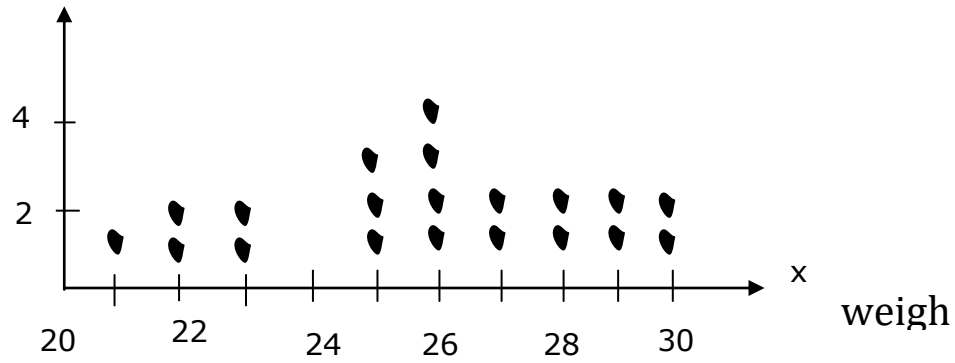


Figure 3.1 Dot plot of weights of children

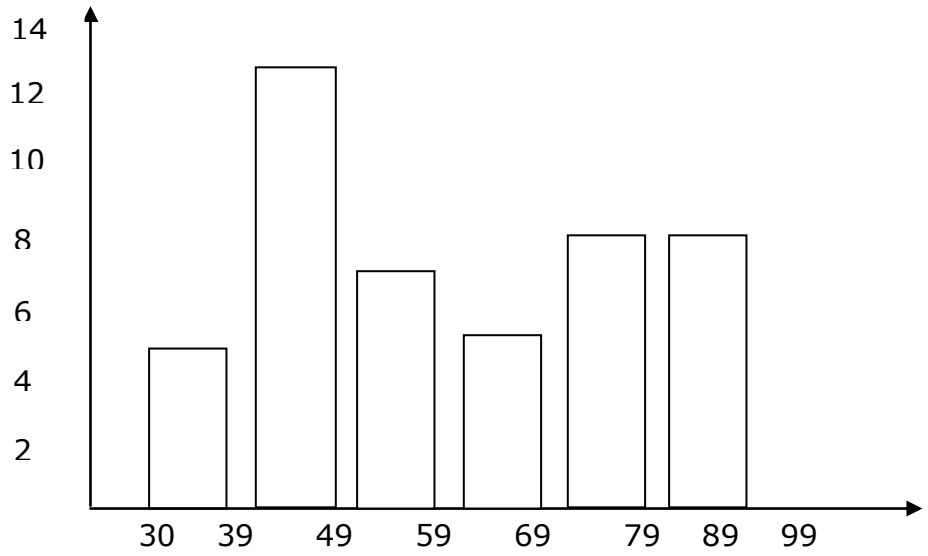
Bar Chart

Bar chart/diagrams are used to represent and compare the frequency distribution of discrete variables and attributes or categorical series. When we represent data using bar diagram, all the bars must have equal width and the distance between bars must be equal. There are different types of bar diagrams, the most important ones are:

Simple bar chart

It is a one-dimensional diagram in which the bar represents the whole of the magnitude. The height or length of each bar indicates the size (frequency) of the figure represented.

Fig. 3.2: Simple bar chart



Multiple bar chart

In this type of chart, the component figures are shown as separate bars adjoining each other. The height of each bar represents the actual value of the component figure. It depicts distributional pattern of more than one variable.

Example

The following table shows the intake through JAMB by the Faculty of Science of a certain University in three consecutive years.

Table 3.4

Department	2002	2003	2004
Botany	43	40	35
Chemistry	28	35	42
Zoology	45	40	35
Computer Science	33	25	28
Physics	40	35	38

Mathematics	35	42	45
Biology	37	40	42
Total	261	257	265

Draw (i) multiple bar chart department by department for the three years.

(ii) a component bar chart.

Solution

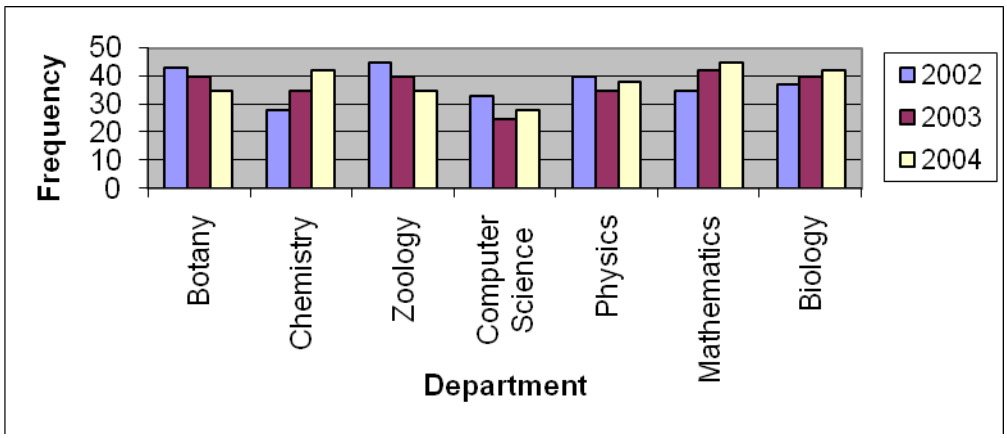


Figure 3.3: Multiple bar chart for JAMB Admission

Component (or sub-divided) Bar Diagram

Bars are sub-divided into component parts of the figure. These sorts of diagrams are constructed when each total is built up from two or more component figures. They can be of two kinds:

Actual Component Bar Diagrams: When the overall height of the bars and the individual component lengths represent actual figures.

Percentage Component Bar Diagram: Where the individual component lengths represent the percentage each component forms the overall total. Note that a series of such bars will all be the same total height, i.e., 100 percent.

(ii)

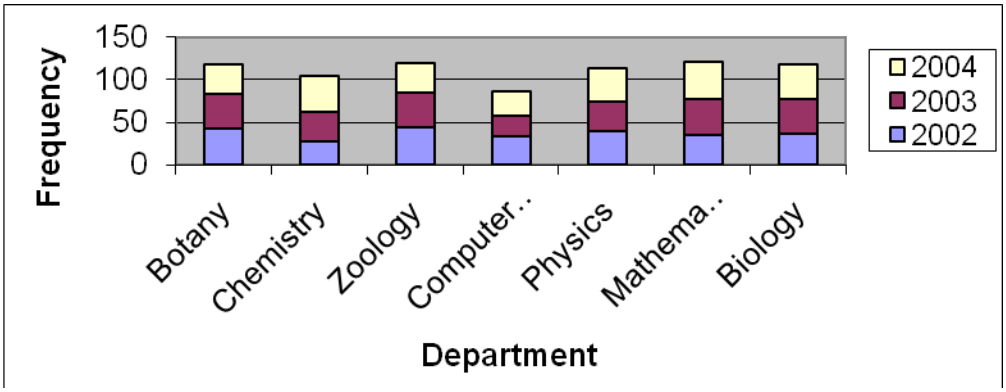


Figure 3.4: Component bar chart for JAMB Admission

Pie-chart

The pie chart (circle graph) is used to display relative frequencies rather than actual frequencies for the data (qualitative or quantitative discrete data). We draw a circle and then divide it into a series of wedges or slices to represent each class in the relative frequency distribution. The size of each slice is proportional to the percentage of the data that fall into the corresponding class.

EXAMPLE

The population of five towns in a country X in 1986 is as follows:

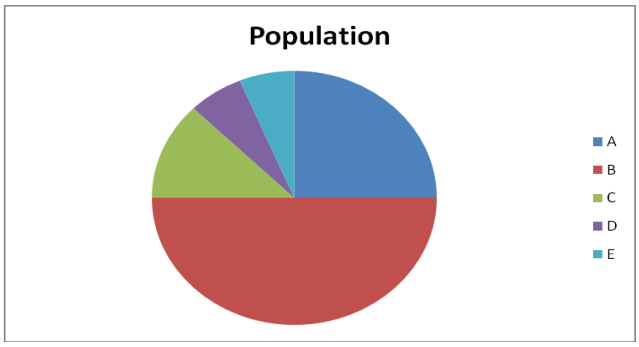
Town	Population
A	50,000
B	100,000
C	25,000
D	12500
E	12500

SOLUTION

Town	Population	Sectorial Angle
A	50,000	90^0
B	100,000	180^0
C	25,000	45^0
D	12500	22.5^0

E	12500	22.5^0

Fig. 3.5: A pie chart showing the population of town X in 1986



Histograms

A histogram is the graph of the frequency distribution of continuous measurement variables (quantitative continuous data). It is constructed on the basis of the following principles:

i. The horizontal axis is a continuous scale running from one extreme end of the distribution to the other. It should be labelled with the name of the variable and the units of measurement.

ii. For each class in the distribution a vertical rectangle is drawn with

(i) its base on the horizontal axis extending from one class boundary of the class to the other class boundary, there will never be any gap between the histogram rectangles.

(ii) the bases of all rectangles will be determined by the width of the class intervals. If a distribution with unequal class-interval is to be presented by means of a histogram, it is necessary to make adjustment for varying magnitudes of the class intervals.

Values for the class boundaries, class limits, or class marks may be labeled along the x-axis. Use whichever one of these sets of class numbers best represents the variable.

Example: use the Table below to draw the histogram.

Table 3.5

Class	Class limits	Frequency	Class boundaries	Class center
1	30 – 39	5	29.5 – 39.5	34.5

2	40 – 49	13	39.5 – 49.5	44.5
3	50 – 59	9	49.5 – 59.5	54.5
4	60 – 69	6	59.5 – 69.5	64.5
5	70 – 79	8	69.5 – 79.5	74.5
6	80 – 89	8	79.5 – 89.5	84.5
7	90 – 99	1	89.5 – 99.5	94.5

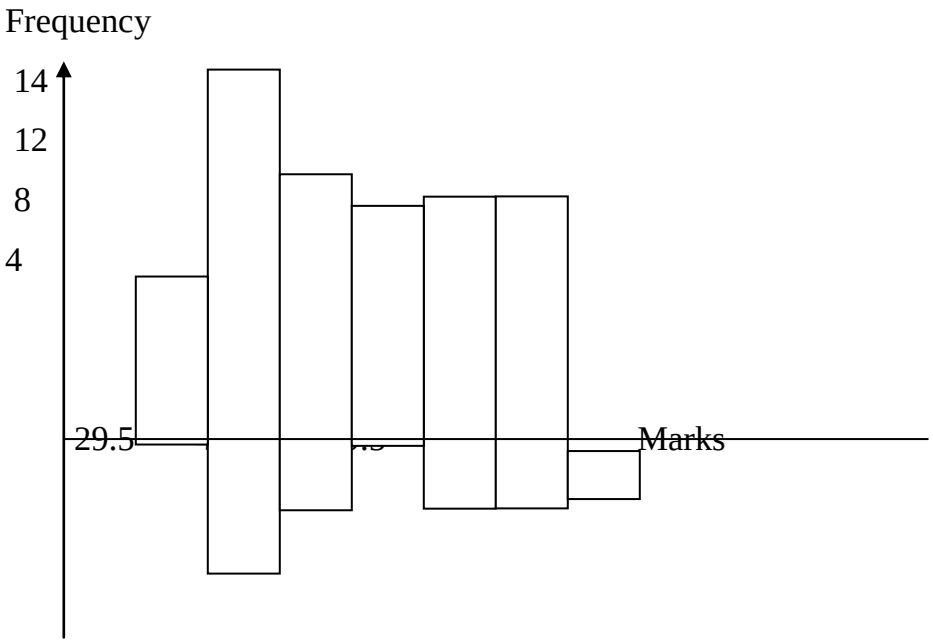


Figure 3.6: Histogram of Marks

Frequency Polygon

If we join the midpoints of the tops of the adjacent rectangles of the histogram with line segments a frequency polygon is obtained. Note that it is not essential to draw histogram in order to obtain frequency polygon. It can be drawn without erecting rectangles of histogram as follows:

i. The scale should be marked in the numerical values of the mid- points of intervals.

ii. Erect ordinates on the midpoints of the interval - the length or altitude of an ordinate representing the frequency of the class on whose mid-point it is erected.
iii. Join the tops of the ordinates and extend the connecting lines to the scale of sizes.

Cumulative Frequency Curve or Ogive

A frequency distribution can easily be converted to a cumulative frequency distribution by replacing the frequencies with cumulative frequencies. When the cumulative frequencies of a distribution are graphed the resulting curve is called Ogive Curve.

The vertical scale represents either the cumulative frequencies or the relative cumulative frequencies. The horizontal scale represents the upper class boundaries. Until the upper class boundary of a class has been reached, you cannot be sure you have accumulated all the data in that class. Therefore, the horizontal scale for an Ogive is always based on the upper class boundaries.

Example

i. Prepare an Ogive from Table 3.5
ii. Give the estimates of the quartiles
iii. Find the median
iv. Estimate the 30 and 70 percentiles
v. Obtain the Range, Interquartile range and semi interquartile range
vi. What number of students scored marks between 60% and 80%?
vii. What will be the pass mark if 60% of the student failed?

Solution

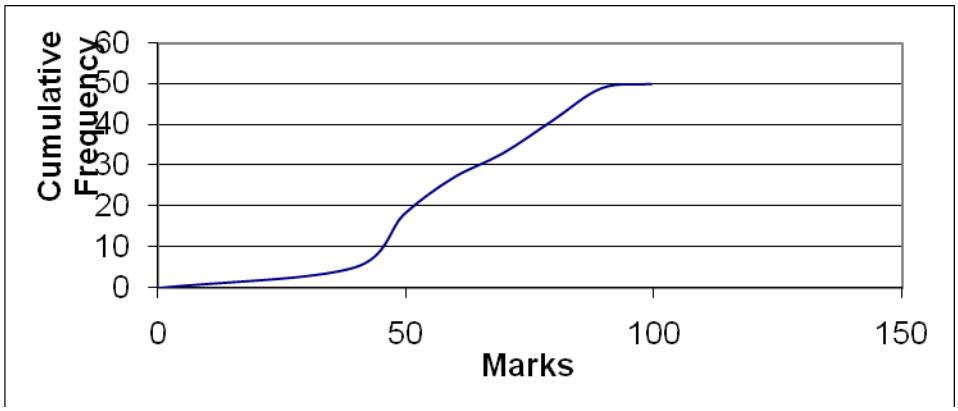


Figure 3.7: Cumulative frequency curve

Quartiles

Q_1 = 25th percentile = 46.5

Q_2 = 50th percentile = 59.5

Q_3 = 75th percentile = 72

median is the 50th percentile and it is equal to 59.5

30th percentile = 49.5

70th percentile = 68

(iii) Range = Highest mark - Lowest mark

= 95 – 31 = 64

from the raw data in section 2.1

Interquartile range = Q_3 - Q_1

= 72 - 46.5 = 25.5

Semi-Interquartile range = $(Q_3 - Q_1)/2$ = (72 - 46.5)/2 = 25.5/2

= 12.75

At 60% mark this intercept the curve at cumulative frequency of 25 students and at 80% mark this intercept the curve at cumulative frequency of 43. Therefore, the number of students that scored between 60% and 80% mark are 43 – 25 = 18 students

(vii) If 60% of the students failed, the pass mark will be from the 60th percentile mark. Trace this to the curve and the pass mark will be 67.

Line graph

Line graphs are diagrammatical representation of the relation between two variables x and y. The co-ordinate points of these variables are joined together to have the line graph.

The line graph is especially useful for the study of some variables according to the passage of time. The time, in weeks, months or years is marked along the horizontal axis; and the value of the quantity that is being studied is marked on the vertical axis. The distance of each plotted point above the base-line indicates its numerical value. The line graph is suitable for depicting a consecutive trend of a series over a long period.

Example

Draw a line graph to represent the information below:

Before	14	20	21	24	22	25	26
After	16	24	23	25	30	27	34

Solution

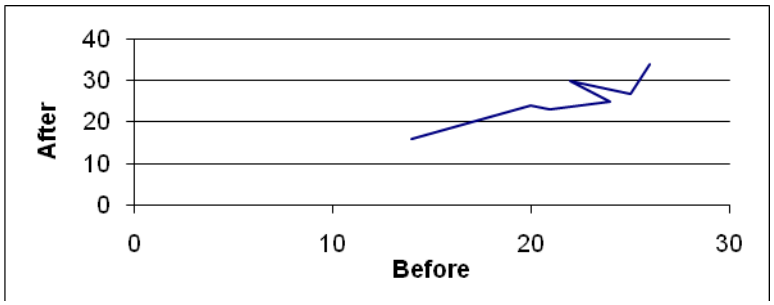


Figure 3.8: Line graph

Box-And-Whisker Diagram

Boxplots are extremely useful graphical devices for describing interval and numerical variables. This plot is based on the five number summary of a set of data (smallest, first quartile, median or second quartile, third quartile and the largest) that is called a box-and-whisker diagram or more simply as a box plot. The three values used (Q_1, Q_2 and Q_3) are sometimes called hinges and is also useful for identifying outliers. The ends of the box

are the lower and upper sample quartiles and the length of the box is the IQR for the variable. The median is marked by a line inside the box. The lines extending from the box extend up to the smallest and largest observation within the interval (Q_1-1.5IQR, Q_3+1.5IQR). Points that are within the interval (Q_1-3IQR, Q_1-1.5IQR) are negative outliers and points that are within the interval (Q_3+1.5IQR, Q_3+3IQR) are positive outliers while those outside the interval (Q_1-3IQR, Q_3+3IQR) are extreme outlies.

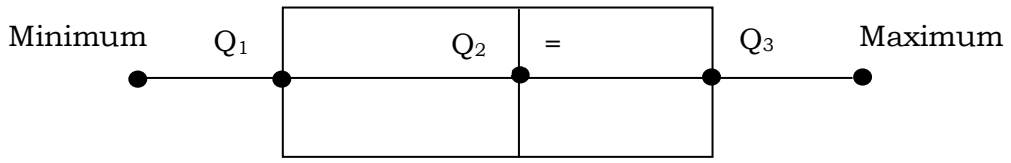


Figure 3.8 Box Plot

To construct a horizontal box-and-whisker diagram, draw a horizontal axis that is scaled to the data. Above the axis draw a rectangle box with the left and right sides drawn at Q_1 and Q_3 with a vertical line segment drawn at the median, Q_2 = median. A left whisker is drawn as a horizontal line segment from the minimum to the midpoint of the left side of the box, and a right whisker is drawn as a horizontal line segment from the midpoint of the right side of the box to the maximum. Note that the length of the box is equal to the interquartile range ($Q_3 - Q_1$). The left and right whiskers contain the first and fourth quarters of the data.

Example

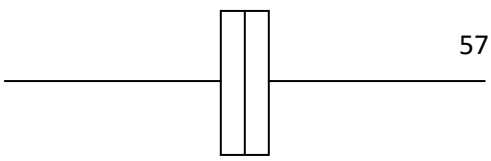
Draw the Box Plot of the data below

24 31 31 40 45 47 48 48 48 49 50 50 50 50 50 50 51 53 53 56 60 70 71 76

Solution

The five number summary are minimum = 24

Q_1 = 47.25, Q_2 = 50, Q_3 = 53, and the maximum = 76



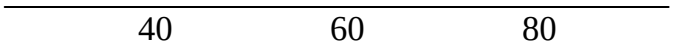


Figure 3.9: Box Plot

4.0 CONCLUSION

Data presentation is a fundamental concept in statistics. Data collected can be in different forms (quantitative or qualitative) and need to be organized and presented. The different methods of presentation have been dealt with in this module.

5.0 SUMMARY

In this module, data exploration was studied further with data organization and presentation using frequency distribution tables and graphical (pictorial) representation of the data

6.0 TUTOR-MARKED ASSIGNMENT

1. The following are baby weighs delivered in a general hospital

2.57	4.21	1.05	3.06	4.50	5.05	3.45	2.15	0.92
3.12	2.67	0.76	4.13	5.93	4.15	2.03	0.57	1.85
4.10	3.41	1.86	2.53	1.46	3.85	5.12	3.24	1.89
2.51	0.95	1.24	2.21	5.86	3.57	2.18	4.29	3.50
0.91	0.82	1.47	4.25	3.81	2.48	1.27	5.35	3.33

Classify these data into a grouped frequency

distribution by using classes of 0.01 – 1.00, 1.01 – 2.00, . . ., 5.01 – 6.00

Find the class width

iii. For the class 4.01 – 5.00, name the value of:

the class center

the class limit,

the class boundaries

iv. Construct a relative frequency histogram of these data.

2**.** The following table shows the frequency distribution of marks of 200 students in a mathematics examination.

Mark	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
Frequency	4	26	52	16	36	34	20	12

i. Draw a cumulative frequency curve and estimate the Quartiles.

ii. Calculate the interquartile and semi-interquartile range from your graph.

iii. Find the pass mark if only 20% of the students should pass.

iv. How many of the students scored between 60% and 85%?

3. Use the table in Exercise 2. to answer the following questions

i. Draw a bar chart of the frequency distribution

ii. Draw a pie chart of the frequency distribution

ii. Draw a histogram of the frequency distribution

4. The following table shows the number admitted into the postgraduate programme for 2 years.

Department	**2003**	**2004**
Chemical Engn.	41	37
Electrical Engn.	40	48
Surveying	35	25
Geography	45	48
Mechanical Engn.	50	45
Geology	35	45

Draw a component and multiple bar charts for this data

5. The year, x, and the birthrate, y, for 1980 – 2000 were as follows:

Year (x)	Birthrate (y)
1980	25,004
1981	25,100
1982	24,345
1983	24,850
1984	23,563
1985	23,236
1986	24,450
1987	18,053
1988	19,245
1989	18,348
1990	15,434
1991	13,347
1992	14,111
1993	15,243
1994	16,172
1995	18,815
1996	17,345
1997	16,457
1998	18,413
1999	19,400
18,721	

i. Construct a line graph of these birthrates.

ii. Interpret your output/result

7.0 REFERENCES AND FURTHER READINGS:

Department of Mathematics (2015). A First Course in Statistics. Department of Mathematics, University of Lagos, Akoka-Yaba, Lagos, Nigeria.

Degu G. and Tessema F. (2007). Biostatistics. In collaboration with the Ethiopia Public Health Training Initiative, The Carter Center, the Ethiopia Ministry of Health, and the Ethiopia Ministry of Education.

Indrayan A. (2017). Statistical medicine: An emerging medical specialty. J Postgrad Med. Available from: http://www.jpgmonline.com/text.asp?2017/63/4/252/216438. Volume 63, (4) 252 – 256.

National Library of Medicine. Epidemiology Studies. National Institutes of Health, 9000 Rockville Pike, Bethesda, Maryland 20892U.S. Department of Health and Human Services. https://toxtutor.nlm.nih.gov/05-003.html.

Petrie A. and Sabin C.(2005). Medical Statistics at a Glance. Published by Blackwell Publishing Ltd, USA.

Song, J. W., & Chung, K. C. (2010). Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, *126*(6), 2234–2242. https://doi.org/10.1097/PRS.0b013e3181f44abc

Study Design 101 by Himmelfarb Health Sciences Library. 2011-2019, The Himmelfarb Health Sciences Library.

MODULE 2: SUMMARY MEASURES, PROBABILITY AND Probability DISTRIBUTIONS

Unit 1: Measures of Location, Measures of Partition and Measures of Spread

Unit 2: Permutations, Combinations and Introduction to Probability

Unit 3: Random Variables and Probability Distributions

Unit 1: Measures of Location, Measures of Partition and Measures of Spread

CONTENTS

1.0 INTRODUCTION

Summarizing data is another way of presenting data to condense the data into a more manageable form and therefore provide a better overall picture of the data. This achieved using some statistical measures which include measures of location, measures of partition and measures of variation.

2.0 OBJECTIVES

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

i. Identify the different methods of data summarization
ii. Compute appropriate summary values for a set of data
iii. Appreciate the properties and limitations of summary values

3.0 MAIN CONTENT

3.1 THE SIGMA (Σ) NOTATION

The Greek capital letter sigma (Σ) is used in Mathematics to indicate the summation of a set of addends. Each of these addends must be of the form of the variable following Σ. For example,

Σx means sum the variable x

Σ (x – 3) means sum the set of addends that are 3 less than the values of each x

Consider a sample of data X_1, ……., X_n, where X_1 corresponds to the first sample point and X_n corresponds to the nth sample point.

Suppose n values of a variable are denoted as x_1, x_2, x_3…., x_n then $\sum x_i = x_1+x_2+ x_3 +\ldots x_n$ where the subscript i range from 1 up to n

Example

Let $x_1=2$, $x_2 = 5$, $x_3=1$, $x_4 =4$, $x_5=10$, $x_6= -5$, $x_7 = 8$

Since there are 7 observations, i range from 1 up to 7

a. $\sum x_i = 2+5+1+4+10-5+8 = 25$
b. $(\sum x_i)^2 = (25)^2 = 625$
c. $\sum x_i^2 = 4 + 25 + 1 + 16 + 100 + 25 + 64 = 235$

Rules for working with summation

i. $\sum(x_i + y_i) = \sum x_i + \sum y_i$, where the number of x values = the number of y values.

ii. $\sum K x_i = k\sum x_i$, where K is a constant.

iii. $\sum K = nK$, where K is a constant.

3.2 MEASURES OF LOCATION OR CENTRAL TENDENCY

Measures of central tendency are numerical values that tend to locate in some sense the middle of a set of data. The term average is often associated with these measures. Each of the several measures of central tendency can be called the average value. They are the mean, median, and mode.

The Arithmetic Mean or simple Mean

To find the mean, x ($\bar{x}$ read "x bar"), you will add all the values of the variable x and divide by the number of these values, n. We express this in formula form as

$$\text{Sample mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3.1)$$

Example

Suppose a sample consists of birth weights (in grams) of all live born infants born at a private hospital in a city, during a 1-week period. This sample is shown as follow:

3265 3323 2581 2759 3260 3649 2841 3248 3245 3200 3609 3314 3484 3031 2838 3101 4146 2069 3541 2834.

$$\text{Sample mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = (3265 + 3260 + \ldots + 2834)/20$$

$$= 633338/20 = 3166.9\text{g}$$

When the sample data has the form of a frequency distribution, we will need to make a slight adaptation in order to find the mean.

Example

Consider the frequency distribution of Table 4.1.

Table 4.1: ungrouped frequency distribution

X	1	2	3	4	5
F	4	8	5	4	7

To calculate the mean $\bar{x}$ using the above formula; we have

$\Sigma x = 1 + 1 + 1 + 1 + 2 + 2 + \ldots + 2 + 3 + \ldots + 3 + 4 + \ldots + 4 + 7 + \ldots + 7$

$\Sigma x = 4(1) + 8(2) + 5(3) + 4(4) + 7(5)$

$= 86 = \Sigma fx$

Therefore, the mean of a frequency distribution may be found by dividing the sum of the data, Σfx, by the sample size, Σf. We can rewrite the formula for use with a frequency distribution as:

Mean = $(\Sigma x f)/\Sigma f$

Example

Table 4.2

x	f	xf
1	4	4
2	8	16
3	5	15
4	4	16
Total	28	86

Mean = 86/28 = 3.07

The arithmetic mean is, in general, a very natural measure of central location. One of its principal limitations, however, is that it is overly sensitive to extreme values. In this instance it may not be representative of the location of the great majority of the sample points.

Harmonic Mean

This is the reciprocal of the average of reciprocals. It is usually represented by x_H and defined by

$$\left[\frac{1}{N}\sum_{j=1}^{n}\frac{1}{X_j}\right]^{-1} = \frac{1}{\frac{1}{N}\sum_{j=1}^{n}\frac{1}{X_j}} = \frac{N}{\sum_{j=1}^{n}\frac{1}{X_j}}.$$

Example

Find the Harmonic mean for the following data 2, 5, 3, 6, 7.

Solution

$x_H = 5/(1/2 + 1/5 + 1/3 + 1/6 + 1/7) = 5/(0.5 + 0.2 + 0.33 + 0.167+ 0.143) = 3.73$

Geometric Mean

This is the nth root of the product of the n numbers in a data set. This is usually represented by x_G and defined by

$$\bar{x}_G = \sqrt[n]{X_1 x X_2 x \ldots x X_n} = (X_1 x X_2 x \ldots x X_n)^{1/n} \qquad (3.6)$$

Example

Find the Geometric mean for the data above in Example 3.6

$\bar{x}_G = \sqrt[5]{2x5x3x6x7} = \sqrt[5]{1260} = 4.17$

Median

An alternative measure of central location, perhaps second in popularity to the arithmetic mean, is the median.

Suppose there are n observations in a sample. If these observations are ordered from smallest to largest, then the median is defined as follows:

The sample median is

i. The $[(n + 1)/2]^{th}$ observations if n is odd

ii. The average of the $[n/2]^{th}$ and $[n/2 +1]^{th}$ observations if n is even.

The median is defined differently when n is even and odd because it is impossible to achieve this goal with one uniform definition. For samples with an odd sample size, there is a unique central point; for example, for sample of size 7, the fourth largest point is the central point in the sense that 3 points are both smaller and larger than it. For samples with an even size, there is no unique central point and the middle 2 values must be averaged. Thus, for sample of size 8, the fourth and the fifth largest points would be averaged to obtain the median, since neither is the central point.

Example

Compute the sample median for the birth weight data

Solution

First arrange the sample in ascending order 2069, 2581, 2759, 2834, 2838, 2841, 3031, 3101, 3200, 3245, 3248, 3260, 3265, 3314, 3323, 3484, 3541, 3609, 3649, 4146

Since n=20 is even,

Median = average of the 10th and 11th largest observation

= (3245 + 3248)/2 = 3246.5 g

Example

Find the median of these numbers 4, 6, 7, 8, 10, 12.

Solution

Depth of the median = $\frac{6 + 1}{2} = 3.5$

This is to say that the median is halfway between the third and fourth pieces of data. To find the number halfway between any two values, add the two values together and divide by 2. In this case, add 7 and 8, then divide by 2. The median is 7.5

The principal strength of the sample median is that it is insensitive to very large or very small values.

Median from a grouped data

The formula for calculating the median from grouped data is defined as

$$\tilde{X} = Lm + \left(\frac{\frac{N}{2} - Cfb}{m} \right) w$$

where Lm = Lower class boundary of the median class

fm = Frequency of median class

N = $\sum f$ is the total frequency

Cfb = Cumulative frequency before the median class

W = Class width.

Mode

The mode for a set of data is the value that occurs most frequently.It is the value of the observation that occurs with the greatest frequency. A particular disadvantage is that,

with a small number of observations, there may be no mode. In addition, sometimes, there may be more than one mode such as when dealing with a bimodal (two-peaks) distribution. It is even less amenable (responsive) to mathematical treatment than the median. The mode is not often used in biological or medical data.

Example

Find the modal values for the following data

(i) 22, 66, 69, 70, 73. (no modal value)

(ii) 1.8, 3.0, 3.3, 2.8, 2.9, 3.6, 3.0, 1.9, 3.2, 3.5 (modal value = 3.0 kg)

Example

Find the modes of the following data. 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4

Solution

The values with the highest number of occurrence are 2 and 4. They both have equal frequency of 4. That is, we have a bimodal case.

Mode from Grouped Data

The mode of a grouped distribution can be obtained either

- ❖ from the frequency curve by finding the value at the highest point or
- ❖ by calculation using the following formula.

From a grouped data the mode is defined as

$$\hat{X} = Lm + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) W$$

where Lm = lower class boundary of the modal class.

Δ_1 = difference between the frequency of the modal class and the class before it.

Δ_2 = difference between the frequency of the modal class and that above it.

w = is the class width.

EXAMPLE FOR GROUPED DATA

The table below presents the age of patients that attended various clinics in the hospital. Determine the mean age of the patients.

Group	Frequency
1 – 5	2
6 – 10	4

11 – 15	8
16 – 20	5
21 – 25	3
26 - 30	1
Total	**23**

SOLUTION

Find the class mark for each class by adding the lower and upper class limits of the class and divide by 2.

Group	**Class Mark (X)**	**f**	**fx**
1 – 5	3	2	6
6 – 10	8	4	32
11 – 15	13	8	104
16 – 20	18	5	90
21 – 25	23	3	69
26 - 30	28	1	28
		23	**329**

$N = \sum f$

$\bar{X} = \frac{\sum fx}{N} = \frac{329}{23} = 14.304$

Mean age = 14.304 years

MEDIAN

Example

The table below shows the number of time that 100 doctors attend to patients in a month at a particular hospital

attendance	**number of doctors (f)**
1 – 2	1
3 – 4	8
5 – 6	26
7 – 8	38

9 – 10	19
11 – 12	7
13 – 14	1

Calculate the median and mode for the number of attendance in the hospital

SOLUTION

The first thing to do is to obtain the cumulative frequency distribution as follow

attendance	**f**	**Cumulative Frequency (cf)**
1 – 2	1	1
3 – 4	8	9
5 - 6	26	35
7 – 8	38	73
9 – 10	19	92
11 -12	7	99
13 -14	1	100

determine $\frac{N}{2} = \frac{100}{2} = 50$, clearly the median value belong to the class

median class = (7 – 8).

Class boundary of the median class = 6.5 – 8.5

lower class boundary (Lm) of the median class is 6.5.

frequency of the median class (fm) is 38

the cumulative frequency before the median class (Cfb) is 35

the class interval (w) is 2 and the median is obtained as

$$\tilde{X} = Lm + \left(\frac{\frac{N}{2} - Cfb}{m} \right) w$$

$$\tilde{X} = 6.5 + \left[\frac{50 - 35}{38} \right] 2$$

$$= 6.5 + 0.789$$

$$= 7.289$$

$$\cong 7$$

MODE

Mode = $\hat{X} = Lm + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) W$

Modal class is the class with highest frequency = 7 – 8
Class boundary for the modal class is 6.5 – 7.5
Mode = 6.5 +$(\frac{38-26}{(38-26)+(38-19)})2$ = 6.5 + 0.77 = 7.27

Modal number of attendance = 7

3.2.1 Properties of Some Statistical Measures

The process of computing any one of the averages discussed so far is comparatively simple. But it is not always easy to choose one particular average which may represent a statistical distribution for the purpose of the inquiry that we have in hand. Below is given a summary of the characteristics, advantages and disadvantages of each average in order to enlarge the awareness of the user so that the selection process could be facilitated.

Mean

Characteristics

i. The value of the arithmetic mean is determined by every item in the series.
ii. It is greatly affected by extreme values.
iii. The sum of the deviations about it is zero. 4) The sum of the squares of deviations from the arithmetic mean is less than of those computed from any other point.

Advantages

i. It is based on all values given in the distribution.
ii. It is most early understood. 3) It is most amenable to algebraic treatment.

Disadvantages

i. It may be greatly affected by extreme items and its usefulness as a “Summary of the whole” may be considerably reduced.
ii. When the distribution has open-end classes, its computation would be based assumption, and therefore may not be valid.

Median

Characteristics

i. It is an average of position.
ii. It is affected by the number of items than by extreme values.

Advantages

i. It is easily calculated and is not much disturbed by extreme values
ii. It is more typical of the series
iii. The median may be located even when the data are incomplete, e.g, when the class intervals are irregular and the final classes have open ends.

Disadvantages

i. The median is not so well suited to algebraic treatment as the arithmetic, geometric and harmonic means.
ii. It is not so generally familiar as the arithmetic mean

Mode

Characteristics

i. It is an average of position
ii. It is not affected by extreme values
iii. It is the most typical value of the distribution

Advantages

i. Since it is the most typical value it is the most descriptive average
ii. Since the mode is usually an "actual value", it indicates the precise value of an important part of the series.

Disadvantages

i. Unless the number of items is fairly large and the distribution reveals a distinct central tendency, the mode has no significance
ii. It is not capable of mathematical treatment
iii. In a small number of items, the mode may not exist.

Geometric Mean

Characteristics

i. It is a calculated value and depends upon the size of all the items.
ii. It gives less importance to extreme items than does the arithmetic mean.

iii. For any series of items, it is always smaller than the arithmetic mean.

iv. It exists ordinarily only for positive values.

Advantages

i. Since it is less affected by extremes it is a more preferable average than the arithmetic mean

ii. It is capable of algebraic treatment

iii. It based on all values given in the distribution.

Disadvantages

i. Its computation is relatively difficult.

ii. It cannot be determined if there is any negative value in the distribution, or where one of the items has a zero value.

3.3 MEASURES OF PARTITION

These are statistical measures used to divide a data set into smaller equal parts. They include: Quartiles, Deciles and Percentiles

Quartiles

They divide data into four parts using three quartiles Q_1, Q_2 and Q_3. The first quartile, Q_1, is referred to as the lower quartile and the last quartile, Q_3, is known as the upper quartile. Q_1 splits the data into the lower 25% of the values and the upper 75% of the values. Similarly, the upper quartile subdivides the data into the lower 75% of the values and the upper 25%. The difference between the upper quartile and the lower quartile is known as the inter-quartile range**,** which indicates the spread of the middle 50% of the data.Basically, there are three types of quartiles, first quartile, second quartile, and third quartile. The other name for the first quartile is lower quartile. The representation of the first quartile is 'Q_1'. The other name for the second quartile is median. The representation of the second quartile is by 'Q_2'. The other name for the third quartile is the upper quartile. The representation of the third quartile is by 'Q_3'.

Example

Find the quartiles of the following numbers:

10 72 18 45 32 56 64 27 60

Solution

Arranging the numbers in ascending order of magnitude,

10 18 27 32 45 56 60 64 72

n = 9,

Depth of $Q_1 = (n+1)/4 = (9 + 1)/4 = 2.5^{th}$

Depth of $Q_2 = 2(n + 1)/4 = 2(9 + 1)/4 = 5^{th}$

Depth of $Q_3 = 3(n + 1)/4 = 3(9 + 1)/4 = 7.5^{th}$

Therefore, $Q_1 = (18 + 27)/2 = 22.5$

$Q_2 = 45$ and $Q_3 = (60 + 64)/2 = 62$.

Deciles

The deciles subdivide data intoten equal parts using D_1, D_2, . . ., D_9. As an example, the fourth decile splits the data into the lower 40% of the values and the upper 60% of the values.Deciles are those values that divide any set of a given observation into a total of ten equal parts. Therefore, there are a total of nine deciles. These representations of these deciles are as follows – D_1, D_2, D_3, D_4, …… D_9.

D_1 is the typical peak value for which one-tenth (1/10) of any given observation is either less or equal to D_1. However, the remaining nine-tenths (9/10) of the same observation is either greater than or equal to the value of D_1.

Percentile

The other name for percentiles is centiles. A centile or a percentile basically divide any given observation into a total of 100 equal parts. The representation of these percentiles or centiles is given as – P_1, P_2, P_3, P_4, …… P_{99}.

P_1 is the typical peak value for which one-hundredth (1/100) of any given observation is either less or equal to P_1. However, the remaining ninety-nine-hundredth (99/100) of the same observation is either greater than or equal to the value of P_1. This takes place once all the given observations are arranged in a specific manner i.e. ascending order.

The 25^{th}, 50^{th}, and 75^{th} percentiles are the first, second, and third quartiles of the sample, denoted as Q_1, Q_2, and Q_3, respectively. The 10^{th}, 20^{th}, 30^{th}, ……, 90^{th} percentiles are the

deciles of the sample. So note that the 50th percentile is also the median, the second quartile, and the fifth deciles.

EXAMPLES

UNGROUP DATA

Locate the position of 1st quartile, 5th decile, 75th percentile, using these medical data below

5, 7, 9, 3, 8, 15, 18, 0, 3, 6, 7, 34, 21, 56, 8, 9, 4, 6, 8, 33

SOLUTION

Re-arrange the data from smallest to highest: 0, 3, 3, 4, 5, 6, 6, 7, 7, 8, 8, 8, 9, 9, 15, 18, 21, 33, 34, 56

1ST QUARTILE $= Q_1 = \frac{N\,X\,1}{4}th = \frac{20\,X\,1}{4}th = 5th$

Therefore, $Q_1 = 5$ (number in the 5th position)

5thDecile $= D_5 = \frac{N\,X\,5}{10}th = \frac{20\,X\,5}{10}th = 10th$

Therefore, $D_5 = 8$ (number in the 10th position)

75thPercentile $= P_{75} = \frac{N\,X\,75}{100}th = \frac{20\,X\,75}{100}th = 15th$

Therefore, $7_{75} = 15$ (number in the 15th position)

GROUPED DATA

For the grouped data, the formula for quartile, decile and percentile is:

$$X = L + \left[\frac{\frac{NXi}{k} - cfb}{f}\right]w$$

Use the data below to find 3rd quartile, 5th decile and 25th percentile

Class	f
10-14	1
15-19	4

20-24	8
25-29	19
30-34	35
35-39	20
40-44	7
45-49	5
50-54	1

SOLUTION

Class	**f**	**cf**
10-14	1	1
15-19	4	5
20-24	8	13
25-29	19	32
30-34	35	67
35-39	20	87
40-44	7	94
45-49	5	99
50-54	1	100

Q_3 position = $\frac{N\,X\,i}{k}th = \frac{100\,X\,3}{4} = 75th$

Class interval that contain $Q_3 = 35 - 39$

Class boundary that contain Q_3 =34.5 -39.5

$$Q_3 = 34.5 + \left\{\frac{\frac{100\,X\,3}{4} - 67}{20}\right\} 5 = 34.5 + 2 = 36.5$$

$Q_3 = 36.5$

D_5 position = $\frac{N\,X\,i}{k}th = \frac{100\,X\,5}{10} = 50th$

Class interval that contain $D_5 = 30 - 34$

Class boundary that contain D_5=29.5 -34.5

$$D_5 = 29.5 + \left\{\frac{\frac{100\,X\,5}{10} - 32}{35}\right\}5 = 29.5 + 2.57 = 32.07$$

$D_5 = 32.07$

P_{25} position = $\frac{N\,X\,i}{k}th = \frac{100\,X\,25}{100} = 25th$

Class interval that contain $P_{25} = 25 - 29$

Class boundary that contain P_{25}=24.5 -29.5

$$P_{25} = 24.5 + \left\{\frac{\frac{100\,X\,25}{100} - 19}{13}\right\}5 = 24.5 + 2.31 = 26.81$$

$P_{25} = 26.81$

3.4 MEASURES OF SPREAD OR VARIATION

It is not enough just to report a number that describes the centre of sample. The spread in a sample is also an important characteristic of a sample. Once the middle of a set of data has been determined, our search for information immediately turns to the measures of dispersion (spread).

These numerical values describe the amount of spread, or variability, that is found among the data.

Consider the following data sets:		Mean
Set 1:	60 40 30 50 60 40 70	50
Set 2:	50 49 49 51 48 50 53	50

The two data sets given above have a mean of 50, but obviously set 1 is more "spread out" than set 2. How do we express this numerically? The object of measuring this scatter or dispersion is to obtain a single summary figure which adequately exhibits whether the distribution is compact or spread out.

The measures of dispersion include the range, interquartile range, variance, standard deviation and coefficient of variation.

Range

The range is defined as the difference between the highest and smallest observation in the data. It is the crudest measure of dispersion. The range is a measure of absolute dispersion and as such cannot be usefully employed for comparing the variability of two distributions expressed in different units.

Range = $X_{max} - X_{min}$

Where, x_{max} = highest (maximum) value in the given distribution. X_{min} = lowest (minimum) value in the given distribution.

In the example given above (the two data sets)
The range of data in set 1 is 70-30 =40
The range of data in set 2 is 53-48 =5
Characteristics

i. Since it is based upon two extreme cases in the entire distribution, the range may be considerably changed if either of the extreme cases happens to drop out, while the removal of any other case would not affect it at all.
ii. It wastes information for it takes no account of the entire data.
iii. The extremes values may be unreliable; that is, they are the most likely to be faulty
iv. Not suitable with regard to the mathematical treatment required in driving the techniques of statistical inference.

Variance and Standard Deviation

Variance is a measure of the spread of the original values about the mean. When we are concerned with a population, the variance is written in terms of the Greek letter and is denoted by σ^{-2} (sigma square)

The variance is a very useful measure of variability because it uses the information provided by every observation in the sample and also it is very easy to handle mathematically. Its main disadvantage is that the units of variance are the square of the units of the original observations.

However, a more useful measure of the spread or variability in a set of data is the standard deviation, which is defined as the square root of the variance.

Standard Deviation (SD) = $\sqrt{Variance}$

Since the standard deviation is the square root of the variance σ^2, the standard deviation is denoted by σ and is found from the formula.

Population standard deviation σ^- = $\sqrt{\frac{\sum(x-\mu)^2}{N}}$

OR $\sqrt{\frac{N(\sum x^2)-(\sum x)^2}{N^2}}$

One special advantage of working with the standard deviation is that it is measured in the same units as the original data. Thus, if the original set of numbers represent weights of a certain type of item, then both the mean and standard deviation are measured in weights.

Sample standard deviation (s) = $\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

or $\sqrt{\frac{n(\sum x^2)-(\sum x)^2}{n(n-1)}}$

That is, instead of dividing by n data points, we divide by n-1. Just as σ^{-2} and σ^- represent the variance and standard deviation of a population, respectively, we use the symbols s^2 and s to stand for the variance and standard deviation, respectively, of a sample.

Coefficient of Variation (C.V)

This is a dimensionless quantity that measures the relative variation between two series observed in different units. Comparison of two distributions with different means and unit of measurement is done using the coefficient of variation.

It is defined as the ratio of the standard deviation and the mean of a set of data expressed as a percentage.

$$C.V = \frac{S}{\overline{X}} \text{ x } 100$$

The distribution with smaller C.V is said to be better

Examples on Measures of Dispersion

UNGROUP DATA

Below is the average of 10 Heads of household randomly selected from a community for Covid-19 laboratory test: 54, 59, 35, 41, 46, 25, 47, 60, 54, 46.

Find the (i) Range (ii) Mean deviation from the mean (iii) Mean deviation from the median (iv) variance (v) standard deviation (vi) coefficient of variation.

SOLUTION

i. Range = $X_{(max)} - X_{(min.)}$

Range = 60 – 25 = 35

ii. Mean Deviation from mean= $MD_{\overline{X}} = \dfrac{\sum|X - \overline{X}|}{n}$

Mean $= \overline{X} = \dfrac{\sum X}{n} = \dfrac{54 + 59 + ... + 46}{10}$

$= 46.7$

$MD_{\overline{X}} = \dfrac{\sum|X - \overline{X}|}{n} = \dfrac{|54 - 46.7| + |59 - 46.7| + ... + |46 - 46.7|}{10}$

(7.3 + 12.3 + 11.7 + 5.7 + 0.7 + 21.7 + 0.3 + 13.3 + 7.3 + 0.7)/10

$= \dfrac{81}{10} = 8.10$

iii. Mean Deviation from median = $MD_{(\hat{X})}$

Array: 25, 35, 41, 46, 46, 47, 54, 54, 59, 60

Median $= \dfrac{X\left(\frac{n}{2}\right) + X\left(\frac{n}{2}+1\right)}{2} = 46.5$

Mean Deviation from median = $MD_{(\hat{X})} = \dfrac{|54 - 46.5| + |59 - 46.5| + - - - - - + |46 - 46.5|}{10}$

$= \dfrac{7.5 + 12.5 + 11.5 + 5.5 + 0.5 + 21.5 + 0.5 + 13.5 + 7.5 + 0.5}{10}$

$= \frac{81}{10}$

$= 8.1$

(iv) Variance $= \frac{\sum(X-\bar{X})^2}{n}$

$= \frac{(54-46.7)^2 + ... + (46-46.7)^2}{10} = 10.87$

(v) Standard Deviation= S $= \sqrt{\frac{\sum(X-\bar{X})^2}{n}}$

$= \sqrt{\frac{(54-46.7)^2 + ... + (46-46.7)^2}{10}}$

(vi) Coefficient of Variation = C.V $= \frac{S}{\bar{X}} \times 100$

$= \frac{10.37}{46.7} \times 100$

$= 22.21$

GROUP DATA

The table below shows the frequency distribution of clinical data

class	Frequency (f)
0 – 10	2
10 – 20	5
20 – 30	8
30 – 40	12
40 – 50	9
50 – 60	5
60 – 70	1

Find the mean deviation from the mean, variance, standard deviation and coefficient of variation for the data.

SOLUTION

Classes	X	f	fx	$X-\bar{X}$	$\lvert X-\bar{X}\rvert$	$f\lvert X-\bar{X}\rvert$	$(X-\bar{X})^2$	$f(X-\bar{X})^2$
0 – 10	5	2	10	- 29.52	29.52	59.04	871.43	1742.86
10 – 20	15	5	75	- 19.52	19.52	97.6	381.03	1905.15
20 – 30	25	8	200	- 9.52	9.52	76.16	90.63	725.04
30 – 40	35	12	420	0.48	0.48	5.76	0.23	2.76
40 – 50	45	9	405	10.48	10.48	94.32	109.83	988.47
50 – 60	35	5	275	20.48	20.48	102.4	419.43	2097.15
60 – 70	65	1	65	30.48	30.48	30.48	929.03	929.03

$$\bar{X}=\frac{\sum fx}{\sum f}=\frac{1450}{42}=34.52$$

i. Mean Deviation from the mean = $\frac{\sum f\lvert X-\bar{X}\rvert}{\sum f}$

$$= \frac{365.76}{42}$$

$$= 11.089$$

ii. Variance= $S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ = (7665.42)/42 = 182.51

iii. Standard deviation = $\sqrt{182.51}$ = 13.51

iv. Coefficient of variation = $C.V=\frac{S}{\bar{X}}$ x 100

$\frac{13.51}{34.52}X\ 100$

= 39.14%

i. Skewness: If extremely low or extremely high observations are present in a distribution, then the mean tends to shift towards those scores. Based on the type

of skewness, (i)Negatively skewed distribution: occurs when majority of scores are at the right end of the curve and a few small scores are scattered at the left end.

ii. Positively skewed distribution: Occurs when the majority of scores are at the left end of the curve and a few extreme large scores are scattered at the right end.

iii. Symmetrical distribution: It is neither positively nor negatively skewed. A curve is symmetrical if one half of the curve is the mirror image of the other half.

In unimodal (one-peak) symmetrical distributions, the mean, median and mode are identical. On the other hand, in unimodal skewed distributions, it is important to remember that the mean, median and mode occur in alphabetical order when the longer tail is at the left of the distribution or in reverse alphabetical order when the longer tail is at the right of the distribution.

4.0 CONCLUSION

Apart from using frequency distribution tables and graphical display for data presentation, data summaries using different measures can be employed to summarized data.

5.0 SUMMARY

In this module, further exploratory analysis was studied using data summary. Statistical measures for summarizing data were identified and described. This include: measures of location, partitions, variation.

6.0 TUTOR-MARKED ASSIGNMENTS

1. Find (a) $\sum x^2$, (b) $(\sum x)^2$, (c) $\sum x \sum y$, (d) $\sum y^2$, (e) $(\sum y)^2$ for the data shown below:

X	3	4	5	6	7
Y	8	9	10	11	

2. The weights, in pounds, of a group of people signing up at a hotel are:

125	141	141	132	155	160	185	165	172	148
131	154	162	148	135	181	172	133	141	135

Find (i) the mean, median and mode of the weights.

3. Two sample brands of bulbs are selected and tested to see how many hours they can be used before running out of use.

Brand A	1134	1157	1811	1858	1958
Brand B	1456	1787	1611	1872	1853

Calculate the mean and standard deviation

Calculate the coefficient of variation

Which of the brands is better?

4. Estimate the mean, standard deviation and median for the following set of data:

Class boundaries	Frequency
151 – 160	50
161 – 170	60
171 – 180	30
181 – 190	35
191 – 200	25
201 – 210	17
211 – 220	13

7.0 REFERENCES AND FURTHER READINGS

Department of Mathematics (2015). A First Course in Statistics. Department of Mathematics, University of Lagos, Akoka-Yaba, Lagos, Nigeria.

Degu G. and Tessema F. (2007). Biostatistics. In collaboration with the Ethiopia Public Health Training Initiative, The Carter Center, the Ethiopia Ministry of Health, and the Ethiopia Ministry of Education.

Indrayan A. (2017). Statistical medicine: An emerging medical specialty. J Postgrad Med. Available from: http://www.jpgmonline.com/text.asp?2017/63/4/252/216438. Volume 63, (4) 252 – 256.

National Library of Medicine. Epidemiology Studies. National Institutes of Health, 9000 Rockville Pike, Bethesda, Maryland 20892U.S. Department of Health and Human Services. https://toxtutor.nlm.nih.gov/05-003.html.

Petrie A. and Sabin C. (2005). Medical Statistics at a Glance. Published by Blackwell Publishing Ltd, USA.

Song, J. W., & Chung, K. C. (2010). Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, *126*(6), 2234–2242. https://doi.org/10.1097/PRS.0b013e3181f44abc

Study Design 101 by Himmelfarb Health Sciences Library. 2011-2019, The Himmelfarb Health Sciences Library.

UNIT 2: PERMUTATIONS, COMBINATIONS AND INTRODUCTION TO PROBABILITY

CONTENTS

6.0 Tutor-Marked Assignments

7.0 References/Further Readings

1.0 INTRODUCTION

Probability is a mathematical language for measuring uncertainty, likelihood or chance. In this module, basic concepts underlying probability theory are introduced. This includes permutations, combinations, definition of basic terms, properties of probability, laws of probability and conditional probability.

2.0 OBJECTIVES

After completing this module, you will be able to:

i. Define and compute the number of ways arranging and selecting items
ii. Define and understand the concepts of Probability
iii. Compute probabilities of events and conditional probabilities

3.0 MAIN CONTENT

3.1 PERMUTATIONS AND COMBINATONS

3.1.1 Permutation

Permutation is a special arrangement of a group of objects in some order. Any other arrangement of the same objects is a different permutation. The key words for permutation are order or arrangement. For example, lets arrange n people in order. There are n possible chances for the first person, n-1 remaining possible chances for the second person, n-2 remaining possible chances for the third person, e.t.c, that is,

The number of possible arrangement = n x (n –1) x (n – 2) x ….x 1)

= n! (n factorial)

Example

0! = 1
1! = 1
2! = 2 x 1 = 2
3! = 3 x 2 x 1 = 6
4! = 4 x 3 x 2 x 1 = 24
5! = 5 x 4 x 3 x 2 x 1 = 120
6! = 6 x 5 x 4 x 3 x 2 x 1 = 720

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

this is the number of permutations of n objects taken r at a time.

Example

In how many ways can three people be seated on 6 seats in a row?

Solution

Arranging 3 people on 6 seats $= {}^{6}P_{3}$

$$6P_3 = \frac{6!}{(6-3)!} = \frac{6!}{3!} = \frac{6 \times 5 \times 4 \times 3!}{3!}$$

$$= 6 \times 5 \times 4 = 120 \text{ ways}$$

Example

How many distinct arrangements can be made using all the letters of the word Economics?

Solution

From the word Economics, o = 2, c = 2, and total letters = 9

$\therefore$ Total arrangement $= \frac{\text{(Number of letter)!}}{\text{(Frequency of letters)!}}$

$$= \frac{9!}{2!\,2!} = \frac{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2}{2 \times 2} = 90720$$

Example

How many different numbers of six digits can be formed using digits 4, 4, 6, 6, 6, 6.

Solution

Total digits (n) = 6

4 has frequency = 2

6 has frequency = 4

Total numbers that can be formed $= \frac{6!}{2!\,4!}$

$$= \frac{6 \times 5 \times 4!}{2 \times 1 \times 4!} = 15$$

3.1.2 Combination

Combination is any collection of a group of objects without regard to order. Problems involving combinations, where order is not relevant, are very similar to problems involving permutations, where order is critical. The only difference between permutations and combinations is whether order matters.

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

is the number of possible combinations of n objects taken r at a time.

Example

Find the number of ways in which three students can be selected from five students.

Solution

3 students can be chosen from 5 students in $^{5}C_{3}$ ways

$= (5!)/ (5 – 3)! 3$

$= 5 \times 2 = 10$ ways

Example

A Mathematics examination consists of 8 questions out of which candidates are to answer 5. In how many ways can each candidate select if

(i) There is no compulsory question

(ii) The first 3 questions are compulsory

(iii) At least 3 out of the first 4 questions are compulsory

Solution

From 8 questions to answer 5 questions, if there is no compulsory question,

we have $^{8}C_{5} = (8!)/ (8-5)! 5!$

$= (8 \times 7 \times 6 \times 5!)/(3 \times 2 \times 5!)$

$= 56$ ways

If the first 3 questions are compulsory, then a candidate can choose 2 more questions from the remaining 5,

$^{5}C_{2} = (5!)/(5-2)! 2!)$

$= 10$ ways

At least 3 out of the first 4 questions are compulsory means the candidate may answer 3 out of the first 4 compulsory questions and 2 from the remaining 4 questions or all the 4 first compulsory questions and 1 from the remaining 4 questions.

$4C_{3} \times 4C_{2} + 4C_{4} \times 4C_{1}$

$4!/(4-3)!\,3! \times 4!/(4-2)!\,2! + 4!/(4-4)!\,4! \times 4!/(4-1)!\,1!$

$= 4 \times 6 + 1 \times 4 = 24 + 4 = 28$ ways

Example

From a gathering of 100 people of which 40 are men, a committee of 15 is to be formed. In how many ways can this be done so that

(i) 3 men are there? (ii) no man is included?

Solution

Total number of people = 100

Men = 40, Women = 100 – 40 = 60

3 men in the committee means 12 women in the committee

$^{40}C_3 \times {}^{60}C_{12} = 40!/(40-3)!\,3! \times 60!/(60-12)!\,12!$

$$= \frac{40!}{37!\,3!} \times \frac{60!}{48!\,12!} = \frac{40 \times 39 \times 38}{6} \times \frac{60!}{48!\,12!}$$

$$= \frac{40 \times 13 \times 19}{48! + 2!} \times \frac{60!}{48!\,12!} = 9880 \times 60!$$

If no man is included, it means the whole of the committee members are women. We have 60 women in the gathering.

$60C_{15} \times 40C_0 = 60C_{15} = \frac{60!}{(60-15)!\,15!} = \frac{60!}{45!\,15!}$

3.2 INTRODUCTION TO PROBABILITY

In statistics, words like '' likelihoods, Chance and uncertainty'' can be used in place of Probability. For example, the chance of an accident occurring on a road, the likelihood of getting a head when a coin is tossed, chance of a top politician winning an election, assessing the degree of uncertainty, in any given situation.

Definition of terms

Experiment: Any well-defined process that yields a result or an observation. Examples are (i) tossing a coin (ii) throwing a die.

Random Experiment: An experiment whose outcome cannot be predetermined.

Outcome: A particular result of an experiment

Sample space (S): The set of all possible outcomes of a random experiment.

Sample point: The individual outcomes in a sample space.

Event: Any subset of the sample space. If A is an event, then n(A) is the number of sample points that belong to event A.

Probability of an event is a measure of the likelihood of that event occurring. If an experiment has a finite number of outcomes which are equally likely, then the probability that an event A will occur is given by

P(A) = (number of outcomes in A)/(Total number of possible outcomes in S)

Some Examples on Probability

Examples

1. A die is tossed once and the outcome could be any of these: The sample space is
   S = { 1, 2, 3, 4, 5, 6 }
2. A coin toss twice or two coins is tossed, the sample space is:
   S = { HH, HT, TH, TT }
3. A coin toss three times or three coins is tossed, the sample space is:
   S = { HHH, HHT, HTH, THH,HTT,THT,TTH,TTT }
4. Two dice are rolled and the sum of the numbers appearing are observed.

Table 5.1

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

with a total of 36-point sample space.

3.2.1 Properties of Probability

A probability is always a numerical value between zero and one.

i. $0 \leq P(A) \leq 1$

ii. (ii) $\sum_{alloutcomes} P(x) = P(S) = 1$

iii. States that if we add up the probabilities of each of the sample points in the sample space, the total probability must equal one.

iv. (iii) $P(\varphi) = 0$. That is, probability of an impossible event is zero.

a. P (A) = 1. The probability of a certain event is one.

v. (v) $P(A') = 1 - P(A)$. The probability that a particular event A will not occur is equal to 1 minus the probability that the event will occur.

Example

Find the probability that a head will appear when two coins are tossed.

Solution

The sample space = {HH, HT, TH, TT}

Let event A be the occurrence of one head.

P (a head will appear) = (numbers of A in sample space)/(number in sample space)

= 2/4 = 0.5

Example

Two dice are rolled and the sum of the numbers appearing are observed. Find the possibility of getting (i) a total of 5 (ii) a total of 12.

Solution

From example 4.3, the sample space is given as

i. P (of getting a total of 5) = $\frac{\text{numbers with total of 5}}{\text{number in sample space}}$

= 4/36

= 1/9

ii. P (of getting a total of 12) = 1/36

3.2.2 Mutually Exclusive Events and The Additive Law

Two events A and B are mutually exclusive if they have no elements in common. If A and B are outcomes of an experiment they cannot both happen at the same time. That is, the occurrence of A precludes the occurrence of B and vice versa. For example, in the toss of a coin, the event A (it lands heads) and event B (it lands tails) are mutually

exclusive. In the throw of a pair of dice, the event A (the sum of faces is 7) and B (the sum of faces is 11) are mutually

The additive law, when applied to two mutually exclusive events, states that the probability of either of the two events occurring is obtained by adding the probabilities of each event. Thus, if A and B are mutually exclusive events,

P(A or B) = P(A) + P(B).

Extension of the additive law to more than two events indicates that if A, B, C… are mutually exclusive events, P(A or B or C or…) = P(A) + P(B)+ P(C) + …

Example

One die is rolled. Sample space = S = (1,2,3,4,5,6).

Let A = the event an odd number turns up, A = (1,3,5)

Let B = the event a 1,2 or 3 turns up; B = (1,2,3)

Let C = the event a 2 turns up, C= (2)

i. Find P(A), P(B) and P(C)

ii. Are A and B; A and C; B and C mutually exclusive? -

Solution

(i) P(A) = P(1) + P(3) + P(5) = 1/6+1/6+ 1/6 = 3/6 = 1/2

P(B) = P(1) + P(2) + P(3) = 1/6+1/6+1/6 = 3/6 = ½

P(C) = P(2) = 1/6

(ii) A and B are not mutually exclusive. Because they have the elements 1 and 3 in common

Similarly, B and C are not mutually exclusive. They have the element 2 in common.

A and C are mutually exclusive. They don't have any element in common.

When A and B are not mutually exclusive P(A or B) = P(A) + P(B) cannot be used. The reason is that in such a situation A and B overlap in a venn diagram, and the elements in the overlap are counted twice. Therefore, when A and B are not mutually exclusive,

P(A or B) = P(A) + P(B) – P(A and B).

The formula considered earlier for mutually exclusive events is a special case of this, since P(A and B) = 0.

Example

Of 200 seniors at a certain college, 98 are women, 34 are majoring in Biology, and 20 Biology majors are women. If one student is chosen at random from the senior class, what is the probability that the choice will be either a Biology major or a women).

Solution

P(Biology major or woman) = P(Biology major) + P(woman) - P (Biology major and woman) =34/200 + 98/200 - 20/200 = 112/200 = 0 .56

3.2.3Independent Events and Multiplication Law

Often there are two events such that the occurrence or nonoccurrence of one does not in any way affect the occurrence or nonoccurrence of the other. This defines independent events. Thus, if events A and B are independent,

P(A and B) = P(A).P(B) OR

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

Example

Independent events situation prevails with the sex of offspring. The chance of a male is approximately ½. Regardless of the sexes of previous offspring, the chance the next child is a male is still ½.

Example

Let two events A and B be defined on the same sample space. Suppose P(B) = 0.2 and P(A ∪ B) = 0.75. Find P(A) such that

A and B are independent

A and B are mutually exclusive

Solution

i. If A and B are independent, then

$P(A \cap B) = P(A)\,P(B)$

Thus $P(A \cup B) = P(A) + P(B) - P(A)\,P(B)$

$= P(A) + 0.2 - P(A) \times 0.2$

$= 0.2 + 0.8\,P(A)$

$0.8\,P(A) = 0.75 - 0.2 = 0.55$

$P(A) = 0.55/0.8$

$= 0.6875$

ii. If A and B are mutually exclusive, then

$P(A \cap B) = 0$

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

$= P(A) + 0.2$

$P(A) = 0.75 - 0.2 = 0.55$

Example

Find the probability of getting:

(i) 2 heads (ii) 1 head (iii) no head if a coin is tossed twice.

Solution

(i)P(getting of head) = ½

P(getting two heads) = P(1st is head) P(2nd is head)

= ½ x ½ = ¼

(ii) P(getting 1 head) = P(1st is head and 2nd is tail) or

P(1st is tail and 2nd is head)

= P(H T) + P(T H)

= ½ x ½ + ½ + ½ = ¼ + ¼ = ½

(iii) P(getting no head) = P(1st is tail and 2nd is tail)

= ½ x ½ = ¼

3.2.3 Conditional Probability

The probability of independent events is straight forward. However, when the events are dependent, then solving them become more complicated.

Conditional Probability written as P(A/B) is the probability of an event A given that a "previous" event B has occurred. The conditional probability of A given B is

$P(A/B) = P(\text{both A and B}) = P(A \cap B)/\ P(B)\ ,\ P(B) \neq 0$

The conditional probability of B given A is

$P(B/A) = P(\text{both A and B}) = P(A \cap B)/\ P(A)\ ,\ P(A) \neq 0$

Example

A coin is tossed thrice. Find the probability that there are two heads (i) given that at least one is a tail (ii) given that the first is a tail.

Solution

Let A: at least one is a tail

B: 2 heads appear

The conditional probability is

P(B/A) = P(2 heads appear/at least one is a tail)

= P(2 heads appear and at least one is a tail)/P(at least one is a tail)

Using the original sample space of all 8 equally likely possible outcomes, we see that

P(at least one is a tail) = 7/8 and

P(2 heads appear and at least one is a tail) = 3/8

Therefore,

$$P(B/A) = P(\text{2 heads appears/at least one is a tail}) = \frac{3/8}{7/8}$$

$$= 3/8 \times 8/7 = 3/7$$

which is the same result as we obtained above.

P(the first is a tail) = 4/8

P(2 heads appear and the first is a tail) = 1/8

Therefore,

P(2 heads appear/the first is a tail)

$$= \frac{P(\text{2 heads appear and the first is a tail})}{P(\text{the first is a tail})}$$

$$= \frac{1/8}{4/8}$$

= 1/8 x 8/4 = ¼

Example

Suppose in country X the chance that an infant lives to age 25 is 0.95, whereas the chance that he lives to age 65 is 0.65. For the latter, it is understood that to survive to age 65 means to survive both from birth to age 25 and from age 25 to 65. What is the chance that a person 25 years of age survives to age 65?

Solution

Notation Event Probability

A	Survive birth to age 25	0.95
A and B	Survive both birth to age 25 and age 25 to 65	0.65
B/A	Survive age 25 to 65 given survival to age 25	?

Then, P(B/A) = P(A n B) / P(A) = .65/.95 =0 .684 .
That is, a person aged 25 has a 68.4 percent chance of living to age 65.

4.0 CONCLUSION

Knowledge of statistics without probability concepts is incomplete. Therefore, elementary probability concepts are being introduced.

5.0 SUMMARY

In this module, Probability was introduced with the study of permutation and combination. Thereafter, basic terms in probability theory were defined. Classical definition of probability given, properties of probability and probability laws described.

6.0 TUTOR-MARKED ASSIGNMENTS

1. How many different signals can be made using 7 flags of different colors on a vertical flagpole if exactly 4 flags are used for each signal?

2. How many five-letter code words are possible using the letters in TANTERLIZER if :

The letters may not be repeated?

The letters may be repeated?

Two unbiased dice are thrown once. What is the probability that:

the sum is 8?

the sum is 8 given that a 3 appears?

at least one 3 is thrown?

3. A fair nine – sided die is rolled once. Let A = {1, 4, 6, 7}, B = {4, 6, 3), C = {8, 9}, D = {2, 5, 7, 8, 9}. Assume that each face has the same probability.

a) Find the values of (i) P(A) (ii) P(B) (iii) P(C) (iv) P(D)

b) Find the values of (i) $P(B \cap D)$ (ii) $P(A \cap C)$, (iii) $P(C \cap D)$

c) Find the values of (i) $P(A \cup C)$, (ii) $P(B \cup D)$, (iii) $P(B \cup C)$

4. Let A and B be independent events with $P(A) = 0.6$ and $P(B) = 0.3$. Compute (a) $P(A \cap B)$, (b) $(A \cup B)$

5. Suppose that $P(A) = 0.45$, $P(B) = 0.6$, and $P(A \cup B) = 0.2$

Find $P(A \cap B)$

Give $P(A/B)$

Give $P(B/A)$

6. A clinical record classifies patients by gender and by type of cancer (A or B). The table is given below.

	Gender	
	Male	Female
Type of Cancer A	57	
B	42	62

If one is selected at random, find the probability that:

the selected individual is a male

the selected individual is a type A

7.0 REFERENCES AND FURTHER READINGS

Department of Mathematics (2015). A First Course in Statistics. Department of Mathematics, University of Lagos, Akoka-Yaba, Lagos, Nigeria.

Degu G. and Tessema F. (2007). Biostatistics. In collaboration with the Ethiopia Public Health Training Initiative, The Carter Center, the Ethiopia Ministry of Health, and the Ethiopia Ministry of Education.

Indrayan A. (2017). Statistical medicine: An emerging medical specialty. J Postgrad Med. Available from: http://www.jpgmonline.com/text.asp?2017/63/4/252/216438. Volume 63, (4) 252 – 256.

National Library of Medicine. Epidemiology Studies. National Institutes of Health, 9000 Rockville Pike, Bethesda, Maryland 20892U.S. Department of Health and Human Services. https://toxtutor.nlm.nih.gov/05-003.html.

Petrie A. and Sabin C.(2005). Medical Statistics at a Glance. Published by Blackwell Publishing Ltd, USA.

Song, J. W., & Chung, K. C. (2010). Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, *126*(6), 2234–2242. https://doi.org/10.1097/PRS.0b013e3181f44abc

Study Design 101 by Himmelfarb Health Sciences Library. 2011-2019, The Himmelfarb Health Sciences Library.

UNIT 3: RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

CONTENTS

7.0 References/Further Readings

1.0 INTRODUCTION

Statistics is concerned with data. The link between sample spaces and data is provided by the concept of a random variable. At a certain point in most probability courses, the sample space is rarely mentioned anymore and we work directly with random variables. In this module, random variables in the form of the Bernoulli, Binomial, Poisson and Normal are studied.

2.0 OBJECTIVES

After completing this module, you will be able to:

(i) Define Random and its properties
(ii) Define Probability functions, Probability distributions and Properties
(iii) Study examples of discrete probability distributions
(iv) Describe an example of continuous probability distribution
(v) Understand the concepts and uses of the standard normal distribution
(vi) Study the relationship between normal distribution and standard normal distribution

3.0 MAIN CONTENT

3.1 RANDOM VARIABLE

A variable that assumes a unique numerical value for each of the outcomes in the sample space of a probability experiment is called a random variable. In other words, a random variable is used to represent the outcome of a probability experiment.

For example, if we toss three coins and the random variable x represents the number of tails that occur, then the only possible values it can assume are x = 0, 1, 2, 3. This is a discrete random variable. (it is called “random” because the value it assumes is the result of a chance, or random event).

There are two types of random variables, the **discrete and continuous** random variable.

A discrete random variable is one that can assume any of a set of possible values that can be counted or listed.

A random variable whose values form a continuum (i.e., have no gaps) such that ranges of values occur with specified probabilities is a continuous random variable.

Another example of discrete random variable is when we roll two dice and observe the sum that appears in both dice. The random variable x will be integer values from 2 to 12. When we deal with a discrete random variable and consider all the possibilities associated with it, we generate a discrete distribution or a probability distribution.

A **probability distribution** (mass function) is a mathematical relationship, or rule, that assigns to any possible value of a discrete random variable X the probability $P(X = x_i)$. This assignment is made for all values xi that have positive probability. The probability distribution can be displayed in the form of a table giving the values and their associated probabilities and/or it can be expressed as a mathematical formula giving the probability of all possible values. Probability Distributions can be classified either as discrete probability distribution or continuous probability distribution. Examples of discrete probability distribution include Bernoulli probability distribution, Binomial probability distribution and Poisson probability distribution. Example of continuous probability distribution include Normal probability distribution

General rules of probability distribution:

(i) The probabilities of a probability distribution must be numbers in the interval from 0 to 1 ($0 \leq P(X=x_i) \leq 1$).

(ii) The sum of all the probabilities for each of the random variable of a probability distribution must be equal to 1 ($\sum P(X=x_i) = 1$).

Example

Toss a coin 3 times. Let x be the number of heads obtained. Find the probability distribution of x.

Solution

Random variable X = xi = 0, 1, 2, 3.

Pr (x = 0) = 1/8 …….. TTT

Pr (x = 1) = 3/8 ……. HTT THT TTH

Pr (x = 2) = 3/8 ……..HHT THH HTH
Pr (x = 3) = 1/8 ……. HHH Probability distribution of X.

X = xi:	9	1	2	3
$P(X=x_i)$	1/8	3/8	3/8	1/8

The required conditions are also satisfied. i) $P(X=x_i) \geq 0$ ii) $\sum P(X=x_i) = 1$.

3.1.1 EXPECTATION AND VARIANCE OF DISCRETE RANDOM VARIABLE

The expectation of a discrete random X, denoted by E(x) or μ, is the expected value or the mean value of the random variable. It is obtained by multiplying each possible value by its respective probability and summing over all the values that have positive probability.

The expected value of a discrete random variable is defined as

$E(X) = \mu = \sum X_i P(X=x_i), \quad i = 1, 2, \ldots, n$

Where the x_i's are the values the random variable assumes with positive probability

The variance of discrete random variable with probability function is given by:

$\sigma^2 = \sum (x - \mu)^2 P(X=x_i)$ or $\sum x^2 P(X=x_i) - \mu^2$

The standard deviation is

$$\sigma = \sqrt{(\sum x^2 P(X=x_i) - \mu^2)}$$

Example

Toss a die once and give the possible outcomes. Find the mean and standard deviation.

Solution

The possible outcomes are 1, 2, 3, 4, 5, or 6, and so the random variable x giving the number on the top face is a discrete random variable. The associated probability distribution is as follows

X	1	2	3	4	5	6
$P(X=x_i)$	1/6	1/6	1/6	1/6	1/6	1/6

To find the mean and standard variance of x:

x	P(x)	x P(x)	x^2	x^2 P(x)
1	1/6	1/6	1	1/6
2	1/6	1/3	4	4/6
3	1/6	½	9	9/6
4	1/6	2/3	16	16/6
5	1/6	5/6	25	25/6
6	1/6	1	36	36/6
Total		7/2		91/6

μ = Mean = $\sum x\,P(x)$ = 7/2 = 3.5

variance = $\sum x^2\,P(x) - \mu^2$ = 91/6 – $(3.5)^2$

= 15.167 - 12.25 ≈ 2.917

standard deviation = $\sqrt{Variance}$ = $\sqrt{2.917}$ = 1.708

Example

The number of calls x to arrive at a switchboard during any 1-minute period is a random variable and has the following probability distribution.

X	0	1	2	3	4
P(X = x)	0.2	0.1	0.3	0.3	0.1

Find the mean and standard deviation of x

Solution

x	p(x)	x p(x)	x^2		x^2 p(x)
0	0.2	0	0		0
1	0.1	0.1	1		0.1
2	0.3	0.6	4		1.2
3	0.3	0.9	9		2.7
4	0.1	0.4	16	1.6	
Total	1.0	2.0			5.6

μ = Mean = $\sum x\,P(x)$ = 2.0

σ^2 = variance = $\sum x^2\,P(x) - \mu^2$ = $5.6 - 2^2$

= 5.6 - 4 = 1.6

σ = S.D. = $\sqrt{1.6}$ = 1.265

3.2 DISCRETE PROBABILITY DISTRIBUTIONS

Bernoulli Distribution

A Bernoulli experiment is a random experiment, the outcome of which can be classified in but one of two mutually exclusive and exhaustive ways, say, success or failure (e.g. true or false, male or female, good or bad, etc). We represent for example, the probability of success, say, p, and failure by 1 – p or q. Bernoulli trial occurs when a Bernoulli experiment is done a number of independent times.

The probability mass function (pmf) is given as

$$f(x) = p^x (1-p)^{1-x} \qquad x = 0, 1 \qquad (5.4)$$

where x is a random variable associated with a Bernoulli trial by defining

x(success) = 1 and x (failure) = 0

We say x has a Bernoulli distribution with parameter p [denoted as X~ Bernoulli (p)]

The expected value or mean of x is

$$E(x) = \sum_{x=0}^{1} x P^x (1-P)^{1-x} = 0(1-p) + (1)(p) = p$$

and the variance of x is

$$\sigma^2 = \text{var}(x) = \sum_{x=0}^{1} (x-p)^2 p^x (1-p)^{1-x}$$

85= $(0-p)^2(1\text{-}p) + (1-p)^2 p$.

= $p(1-p) = pq$

as usual, the standard deviation of X will be

$$\sigma = \sqrt{p(1-p)} = \sqrt{pq}$$

Example 5.3: If X~ Ber (0.7), find the mean and variance.

Solution

It can be deduced that p = 0.7 and q = 1 – p = 1 – 0.7 = 0.3

Therefore,

Mean = $\mu = E(x) = p = 0.7$

Variance = $\sigma^2 = V(x) = pq = p(1\text{-}p) = 0.7 \times 0.3$

= 0.21

Binomial Distribution

If we let the random variable X equal the number of observed successes in n Bernoulli trials, the possible values of x are 0, 1, 2,..., n. If x successes occur, where x = 0, 1, 2, …., n, then n–x failures occur. The pdf of x, say f(x) is

$$f(x) = {}_nC_x P^x (1 - P)^{n-x}, \quad x = 0, 1, 2, \ldots. n \qquad (5.5)$$

where

$${}_nC_x = \frac{n!}{x!\,(n - x)!}$$

The number of ways of selecting x positions for the x successes in the n trials. Where (5.5) is the binomial distribution of the random variable X.

The binomial experiment must possess the following properties:

i. Each trial has two possible outcomes (success, failure)
ii. There are n repeated independent trial
iii. The probability of success on each trial is a constant p; the probability of failure is q = 1 – p.
iv. The random variable X equals the number of successes in the n trials.
v. The random variable x defined above is said to be a Binomial distribution with parameters n and p denoted as X~ Bi (n, p)

Example

If X is a binomial random variable, calculate the probability of x for

i. n = 3, x = 2, P = 0.3
ii. n = 4, x = 0, P = 0.4

Solution

$$P(X = x) = {}_nC_x p^x (1 - p)^{n-x}$$

i. $P(X = 2) = {}_3C_2 (0.3)^2 (1\text{-}0.3)^{3-2}$

$= 3 \times 0.09 \times 0.7 = 0.189$

ii. $P(x = 0) = {}_4C_0 (0.4)^0 (1 - 0.4)^{4-0}$

$= (0.6)^4 = 0.1296$

Example

Suppose that in a certain malarious area past experience indicates that the probability of a person with a high fever will be positive for malaria is 0.7. Consider 3 randomly selected patients (with high fever) in that same area.

i. What is the probability that no patient will be positive for malaria?
ii. What is the probability that exactly one patient will be positive for malaria?
iii. What is the probability that exactly two of the patients will be positive for malaria?
iv. What is the probability that all patients will be positive for malaria?
v. Find the mean and the SD of the probability distribution given above.

Solution: (i) 0.027 (ii) 0.189 (iii) 0.441 (iv) 0.343
(v) $\mu = 2.1$ and $\sigma = 0.794$

Poisson Distribution

Poisson distribution is another discrete probability distribution. Some experiments result in counting the number of times particular events occur in a given time. For example, we could count the number of road accidents that occur on the third mainland bridge in Lagos between 2 and 4pm, the number of phone calls arriving at an office between 5 and 6pm, or the number of defective antenna produced by a company in a day. All these three examples have certain characteristics in common.

First, in each case, we are looking at situations where there are relatively few successes during the indicated time period. Thus, the probability of success in any sufficiently small time interval will be quite small.

Second, the individual successes are independent of one another. That is the fact that one accident occurs a month ago should not influence any other accident from occurring today.

Third, it is assumed that the event occurring is uniformly distributed over the entire time period under consideration. Thus, we assume that a phone call arriving into an office will arrive at any time during the day with equal likelihood.

Under the above conditions, the probabilities involved follow a Poisson probability distribution. The Poisson probability formula giving the probability that x success occur during a given time interval is:

$$P(\text{x successes}) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0, 1, 2,\ldots \qquad (5.6)$$

where λ is the average number of successes occurring in the given time interval and the symbol e represents the number

$$e \approx 2.71828$$

When a random variable x has this distribution we write in shorthand $X \sim P_0(\lambda)$. For the Poisson distribution,

Mean (μ) = variance = λ

Table II in the Appendix gives values of the distribution function of a Poisson random variable for selected values of λ.

Example

Suppose X is a Poisson random variable with parameter 4 [$X \sim P_0(4)$]. Compute the following probabilities.

i. $P(X \leq 4)$
ii. $P(X < 5)$
iii. $P(X = 7)$
iv. $P(X > 6)$
v. $P(X \geq 5)$
vi. $P(2 \leq X \leq 5)$
vii. $P(2 < X \leq 5)$
viii. $P(2 \leq X < 5)$

Solution

We use the cumulative Poisson probability table to solve these problems. Here $\lambda = 4$ and he probability $P(x \leq x)$ for selected values of λ and x are given in statistical table.

i. $P(X \leq 4) = 0.629$

ii. $P(X < 5) = P(X \leq 4) = 0.629$

iii. $P(X = 7) = P(X \leq 7) - P(X \leq 6)$
$= 0.949 - 0.889$
$= 0.060$

iv. $P(X > 6) = 1 - P(X \leq 6)$
$= 1 - 0.889 = 0.111$

v. $P(X \geq 5) = P(X \geq 5) = 1 - P(X \leq 4)$
$= 1 - 0.629 = 0.371$

vi. $P(2 \leq X \leq 5) = P(X \leq 5) - P(X \leq 1)$
$= 0.785 - 0.092$
$= 0.693$

vii. $P(2 < X \leq 5) = P(X \leq 5) - P(X \leq 2)$
$= 0.785 - 0.238 = 0.547$

viii. $P(2 \leq X < 5) = P(X \leq 4) - P(X \leq 1)$
$= 0.629 - 0.092 = 0.537$

3.3 NORMAL DISTRIBUTION

The Normal Distribution, also called Gaussian distribution, is most important probability distribution in statistics. The distributions of many medical measurements in populations follow a normal distribution (e.g. Serum uric acid levels, cholesterol levels, blood pressure, height and weights e.t.c.).

For the normal distribution there are two parameters (μ and σ^2) that define and describe it. The normal distribution is a theoretical, continuous probability distribution whose probability density function (PDF), f(x) is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\frac{(x-\mu)^2}{\sigma}\right] \forall\ x \in \text{set of Real numbers}$$

The area that represents the probability between two points c and d is defined by:

$$P(a \leq x \leq = b) = \int_a^b f(x)dx$$

The important characteristics of the Normal Distribution are:

i. It is a continuous probability distribution with continuous variable.

ii. The random variable extends from minus infinity($-\infty$) to plus infinity ($+\infty$).

iii. It is unimodal, bell-shaped and symmetrical about x = u.

iv. It is determined by two quantities: its mean (μ) and SD (σ). Changing μ alone shifts the entire normal curve to the left or right. Changing σ alone changes the degree to which the distribution is spread out.

v. The height of the frequency curve, which is called the probability density, cannot be taken as the probability of a particular value. This is because for a continuous variable there are infinitely many possible values so that the probability of any specific value is zero.

vi. It is centered at the mean μ

vii. $\int_{-\infty}^{\infty} f(x)\,dx = 1$, that is, the total area under the normal distribution curve is 1.

3.4 THE STANDARD NORMAL DISTRIBUTION

An observation from a normal distribution can be related to a standard normal distribution which has a published table. Since the values of μ and σ will depend on the particular problem in hand and tables of the normal distribution cannot be published for all values of μ and σ, calculations are made by referring to the standard normal distribution which has $\mu = 0$ and $\sigma = 1$. Thus an observation x from a normal distribution with mean μ and standard deviation σ can be related to a Standard normal distribution.

To do this, we use the fact that the number of standard deviations that a given measurement x lies away from the mean μ is precisely the associated Z value defined as

$$\{ x - (\text{mean of } x)\}/(\text{standard deviation of } x)$$

Therefore, we can write the formula for Z as

$$Z = (x - \mu)/\sigma$$

(note that if $x = \mu$, then $Z = 0$)

The Z–score is known as a "standardized" variable because its units are standard deviations. The normal probability distribution associated with this standard score Z is called the standard normal distribution.

To find the area under a normal curve (with mean μ and standard deviation σ) between x=a and x=b, find the Z scores corresponding to a and b (call them Z_1 and Z_2) and then find the area under the standard normal curve between Z_1 and Z_2 from the published table (Statistical Table).

Examples of Probabilities Computation using standard score Z

1. Assume a distribution has a mean of 70 and a standard deviation of 10.

How many standard deviation units above the mean is a score of 80?

$Z = (80-70) / 10 = 1$

How many standard deviation units above the mean is a score of 83?

$Z = (83 - 70) / 10 = 1.3$

2. Find the following probabilities using the statistical tables.
   i. $P(Z < 2.5)$
   ii. $P(Z > - 0.35)$
   iii. $P(-1.5< Z< 2.5)$
   iv. $P(-2.85 < Z < -0.93)$
   v. $P(1.21 < Z < 3.2)$
   vi. $P(0 < Z < 1.45)$
   vii. $P(-1.72 < Z < 0)$

Solution

i. Checking the normal distribution table for the area corresponding to Z = 2.5 and conclude that: $P(Z < 2.5) = 0.9938$

ii. Consulting with the table reveals Z = -0.35 is 0.3632. Since the total area under the normal distribution is 1, we see that the area beyond Z = -0.35 is given as 1 – 0.3632 = 0.6368. i.e. $P(Z > - 0.35) = 0.6368$

iii. The areas under the curve $P(Z < -1.5)$ and $P(Z <2.5)$ are 0.0668 and 0.9938 respectively. The final answer is the difference between these two quantities, so that: $P(-1.5 < Z<2.5) = 0.9938 – 0.0668 = 0.927$

iv. The areas under the curve for $P(Z < -2.85)$ and $P(Z < -0.93)$ are 0.0022 and 0.1762 respectively. The final answer is the difference between these two quantities, so that: $P(-2.85 < Z < -0.93) = 0.1762 - 0.0022 = 0.174$

v. The areas under the curve for $P(Z < 1.21)$ and $P(Z < 3.2)$ are 0.8869 and 0.9993 respectively. The final answer is the difference between these two quantities, so that: $P(1.21 < Z < 3.2) = 0.9993 - 0.8869 = 0.1124$

vi. The area under the curve for $P(Z < 1.45)$ is 0.9265 and the total area on each side of the center in a normal distribution is 0.5. We see that the difference between these two quantities gives the desired probability.

$$P(0 < Z < 1.45) = 0.9265 - 0.5 = 0.4265$$

vii. Similarly, as above, the area under the curve for $P(Z < -1.72)$ is 0.0427 and the total area on each side of the center in a normal distribution is 0.5. Therefore,

i. $P(-1.72 < Z < 0) = 0.5 - 0.0427 = 0.4573$

3. Suppose that total carbohydrate intake in 12-14-year-old males is normally distributed with mean 124 g/1000 cal and SD 20 g/1000 cal.

i. What percent of boys in this age range have carbohydrate intake above 140g/1000 cal?

ii. What percent of boys in this age range have carbohydrate intake below 90g/1000 cal?

Solution

Let X be carbohydrate intake in 12-14-year-old males and $X \sim N(124, 400)$

i. $P(X > 140) = P(Z > (140-124)/20) = P(Z > 0.8)$

$= 1 - P(Z < 0.8) = 1 - 0.7881 = 0.2119$

ii. $P(X < 90) = P(Z < (90-124)/20) = P(Z < -1.7)$

$= P(Z > 1.7) = 1 - P(Z < 1.7) = 1 - 0.9554 = 0.0446$

4.0 CONCLUSION

Further to the concepts of probability introduced in module five, the concept of random variables and probability distributions is being discussed.

5.0 SUMMARY

In this module, basic terms like random variable, probabilities functions and probability distributions were defined with examples. Examples of discrete probability distributions and continuous probability distribution were discussed. Examples of computations of probabilities were given.

6.0 TUTOR-MARKED ASSIGNMENTS

1. It is assumed that 20% of the children in a locality are infected with cholera. Suppose that eleven children are selected at random. Let X equals the number of children that are infected. Find

   i. $P(X \leq 2)$
   ii. $P(X = 4)$
   iii. $P(X \geq 4)$
   iv. $P(X > 7)$
   v. $P(X < 9)$

2. It is believed that 40% of workers do not have any health insurance. Suppose that this is true and let X equal the number with no health insurance in a random sample of n = 25 workers.

Find:

   i. the mean, variance and standard deviation of X
   ii. $P(X \geq 20)$
   iii. $P(X \leq 5)$
   iv. $P(X = 10)$

3. Let X have a Poisson distribution with mean 7. Find

   i. $P(X \geq 6)$

ii. $P(X \leq 3)$

iii. $P(3 \leq X \leq 6)$

4. Let X have a Poisson distribution with a variance of 5. Find $P(X = 4)$

5. The probability of passing a course is 0.6. Let 10 students be picked at random. Find the probability of obtaining

i. no failure

ii. no success

iii. five failures and five successes

6. Find the area under the standard normal curve that corresponds to the following Z values:

i. between –1 and 2.5

ii. greater than 2.15

iii. less than 1.51

7. Find the area under the normal curve that lies between the following pairs of Z values

i. Z = -2.85 to Z = -1.55

ii. Z = 0 to 3.41

iii. Z = -1.86 to 1.51

iv. Z = -0.73 to 0.93

8. Find the following:

i. $P(0 < Z < 3.10)$

ii. $P(-1.87 < Z < -0.96)$

iii. $P(0.79 < Z < 2.87)$

iv. $P(-0.93 < Z < 1.61)$

9. Find the following:

i. $P(Z < 2.3)$

ii. $P(Z < -1.76)$

iii. $P(Z > 0.5)$

iv. $P(Z > -1.56)$

10. Find the area under the standard normal curve that corresponds to the following Z values:

i. between –1 and 2.5

ii. greater than 2.15

iii. less than 1.51

11. Find the area under the normal curve that lies between the following pairs of Z values

i. $Z = -2.85$ to $Z = -1.55$

ii. $Z = 0$ to 3.41

iii. $Z = -1.86$ to 1.51

iv. $Z = -0.73$ to 0.93

12. Find the following:

i. $P(0 < Z < 3.10)$

ii. $P(-1.87 < Z < -0.96)$

iii. $P(0.79 < Z < 2.87)$

iv. $P(-0.93 < Z < 1.61)$

13. Find the following:

i. $P(Z < 2.3)$

ii. $P(Z < -1.76)$

iii. $P(Z > 0.5)$

iv. $P(Z > -1.56)$

7.0 REFERENCES AND FURTHER READINGS

Department of Mathematics (2015). A First Course in Statistics. Department of Mathematics, University of Lagos, Akoka-Yaba, Lagos, Nigeria.

Degu G. and Tessema F. (2007). Biostatistics. In collaboration with the Ethiopia Public Health Training Initiative, The Carter Center, the Ethiopia Ministry of Health, and the Ethiopia Ministry of Education.

Indrayan A. (2017). Statistical medicine: An emerging medical specialty. J Postgrad Med. Available from: http://www.jpgmonline.com/text.asp?2017/63/4/252/216438. Volume 63, (4) 252 – 256.

National Library of Medicine. Epidemiology Studies. National Institutes of Health, 9000 Rockville Pike, Bethesda, Maryland 20892U.S. Department of Health and Human Services. https://toxtutor.nlm.nih.gov/05-003.html.

Petrie A. and Sabin C.(2005). Medical Statistics at a Glance. Published by Blackwell Publishing Ltd, USA.

Song, J. W., & Chung, K. C. (2010). Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, *126*(6), 2234–2242. https://doi.org/10.1097/PRS.0b013e3181f44abc

Study Design 101 by Himmelfarb Health Sciences Library. 2011-2019, The Himmelfarb Health Sciences Library.

MODULE 7 SAMPLING DISTRIBUTION AND ESTIMATION

UNIT 1 SAMPLING DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

Before now you have learnt different probability models for disease situations and health outcomes, in which case it was always assumed that the specific probability distributions were known. For example, it can be assumed that; the number of preterm births out of a group n = 100 deliveries at Lagos University Teaching Hospital (LUTH) follows a binomial distribution with parameter (probability) p = 0.5; the number of accidents recorded on a particular highway per day follows a Poisson distribution with parameter (mean) λ = 5; or that the weights of women in reproductive age are normally distributed with parameters (mean) μ = 68kg and (standard deviation) σ = 10kg.

In each of the above examples, it was assumed that data come from a known probability distribution with known underlying properties (or parameters) and that the knowledge of

these properties can be used to make predictions about the behavior of the data. For instance, given the above distribution of weights of women in reproductive age with μ = 68kg and σ = 10kg, and if data on weights of women in reproductive age are collected, it can be predicted that about 78.8% of all women in reproductive age should have weights greater than 60kg.

Now, supposing we have a data set and we do not have an idea of the properties of the underlying distribution. The major problem will be that of guessing or inferring the characteristics (or parameters) of the underlying distribution from the data set. This process is known as Statistical Inference, whereby the value of a (population) parameter is being inferred from a (sample) data. Inferential Statistics is the process of using the value of a sample statistic to make a guess about the value of a population parameter and draw conclusion based on that guess. The process leads to making informed decision, within reasonable level precision and accuracy. Statistical inference can be subdivided into two main areas: estimation and hypothesis testing. Estimation will be dealt with in this module, alongside sampling distributions.

2.0 OBJECTIVES

At the end of this module, you should be to;

i. Differentiate between a population and a sample

ii. Explain the concept of sampling distribution

3.0 MAIN CONTENT

3.1 REVIEW OF POPULATION AND SAMPLE

To understand the concept of sampling distributions, there is need to review commonly used terms like population and sample.

3.1.1 Population

A population is the aggregate number of subjects or respondents or experimental units under consideration. This implies that a population should contain every member of a group of interest, being studied. Example could include the following: all children aged between five and ten with corona virus living in Lagos; all women in Nigeria who had their first children at the age of 15; all women in reproduction age in Ogun State.

A numerical characteristic or quantity or property of the population that we wish to know about is called a population parameter. Let a population consists all HIV patients between the ages of 20 and 60 in Nigeria; with interest to study their weights. A population parameter in this case could be the average weights of the patients. So, population average or mean (μ) is a parameter. Other examples of population parameters are the population variance (σ^2), population standard deviation (σ), population proportion (P), e.t.c. A parameter of a population is usually unknown.

The target of scientific investigation is usually the population, with some characteristics of interest. Sometimes, the population does not exist, like in prospective studies, or could be so large that obtaining measurements of all the members on the characteristics of interest is impossible. This means that obtaining value for the parameters is not possible. When situation like this arise, a sample from the population is studied instead of the entire population. Another way to deal with this problem is to conduct a clinical trial, if it is health-related and if necessary.

3.1.2 Sample

A sample is the part of a population that we actually study. A numerical characteristic or quantity or property of a sample that we wish to know about is called a sample statistic, or simply statistic. Example of sample statistics are sample average ($\overline{X}$), sample variance (S^2), sample standard deviation (S) and sample proportion (p). For instance, we might decide to select 50 HIV patientsin the age range of 20 to 60 and measure their weights. Suppose the average weight of the 50 patients is, say, 58kg.The sample size is 50 and the value of the sample statistic (average weight in this example) is 58kg.

For the process of statistical inference to be valid we must ensure that we take a representative sample of our population. This implies that the characteristics of a sample we take, as much as possible, match the characteristics of the population we are sampling from. Two simple and effective methods of doing this are making sure the sample size is large enough and making sure it is randomly selected. A large sample size includes more members of the population and is more likely to be representative of a population than a small one. For studying the average weight of HIV patients between 20 to 60 years of age in Nigeria, a sample of size1,000,000 HIV patients is more likely to be a representative sample for the population average than a sample of size 10, because the average weight of 1,000,000 patients would be likely close to the population average than that of 10 patients. Depending on the population size, a sample size greater than 30 is taken as large sample.

A random sample is a sample which gives every member of the population an equal chance of being selected into the sample. A random sample has the advantage of eliminating bias. Besides, most statistical procedures are based on the idea of random sampling.

Descriptive Statistics involve the use of techniques and measures to present, explore and summarize sample data. Sample statistics are summary measures. The main features of a sample and further details of random sampling techniques would have been discussed in earlier modules and well documented in most elementary statistics textbooks. Refer to them before you move on.

3.2 SAMPLING DISTRIBUTION OF A SAMPLE STATISTIC

It is also important to state here that a sample statistic or an estimator is a random variable (refer to early chapters of this material for notes on random variables). This means that repeated samples of same size can be drawn independently from the same population and values of the statistic of interest computed for each sample drawn. This process would produce a set of values for the statistics. Therefore, like any random

variable, a statistic (an estimator) has a distribution, referred to as sampling distribution, with mean and variance. Therefore, a sampling distribution of a statistic is defined as the distribution of the values of the statistic over all possible repeated samples of same size drawn from a specific population in consideration.

A very important property of a statistic is unbiasedness and this can be accessed through its sampling distribution. Bias is the difference between the true value of a parameter and the expected value of the statistic (estimator) used in estimating it.

As an illustration: Let a population consist of 2, 3, 5, 7, 9. Then the population mean, say $\mu = (2+3+5+7+9)/5 = 26/5 = 5.2$. Recall that μ is the parameter. Suppose that random samples each of size 3 are to be selected from this population, and their sample means (sample statistic) be computed. Then we would have ${}^{N}C_{n} = {}^{5}C_{3} = 10$ possible random samples, as follows:

	X	Sample	Statistic ($\bar{X}$)
	X1	(2, 3, 5)	(2+3+5)/3 = 3.33
	X2	(2, 3, 7)	(2+3+7)/3 = 4.00
	X3	(2, 3, 9)	(2+3+9)/3 = 4.67
	X4	(2, 5, 7)	(2+5+7)/3 = 4.67
	X5	(2, 5, 9)	(2+5+9)/3 = 5.33
	X6	(2, 7, 9)	(2+7+9)/3 = 6.00
X7		(3, 5, 7)	(3+5+7)/3 = 5.00
	X8	(3, 5, 9)	(3+5+9)/3 = 5.67
	X9	(3, 7, 9)	(3+7+9)/3 = 6.33
X10		(5, 7, 9)	(5+7+9)/3 = 7.00

It is seen that the sample statistic, in this case, the sample mean ($\bar{X}$) is a random variable with some observed values: 3.33, 4, 4.67, 4.67, 5.33, 6, 5, 5.67, 6.33, 7. So we can obtain the mean of $\bar{X}$ as

(3.33+4+4.67+4.67+5.33+6+5+5.67+6.33+7)/10 = 52/10 = 5.2

The above illustration gives the sampling distribution of the sample mean, $\overline{X}$. It can be seen that the mean of $\overline{X}$ is equal to the population mean μ and this implies that $\overline{X}$ is an unbiased statistic (estimator) for estimating μ. If they are not equal, then the statistic would be a biased statistic for estimating μ and the difference is called bias. The mean of the sampling distribution of a statistic, say *h*, could be referred to as its expected value, denoted by E(*h*). An unbiased statistic is often desirable. Good enough, most of the statistics we will be considering like the sample mean and proportion are unbiased estimators of their corresponding population parameters.

Notice that the number of possible samples was generated without replacement and the number of samples was obtained using ${}^{N}C_{n}$, where *N* = population Size and *n* = sample size. Sample selection can be done with replacement, in which case the number of possible samples would be obtained by using N^n.

Variance, which is a measure of variability, is a very important feature of sampling distribution. A sample statistic with smaller variance is often desirable because it is likely to produce a better estimate of the corresponding parameter. In this way, the variance of the sampling distribution of a statistic can be used as a measure of precision. The square root of this variance is called the standard error (SE) of the statistic. In Statistics, an unbiased statistic is likely to have a smaller variance.

The variance of any sample statistic can be obtained under three conditions. Case 1 is when the population variance is known, case 2 is when the population variance is not known but the sample size is large ($n \geq 30$) and case 3 is when the population variance is not known and the sample size is small ($n < 30$). These three cases determine what the sampling distribution of the test statistic would look like and will be considered under the next section.

The Central Limit Theorem is a theorem in Statistics which states that if the number of possible samples drawn from a population is large enough then, whether the population distribution is normal or non-normal, the sampling distribution of the sample mean will

always be normal. To achieve this one needs to sample from an infinitely large population or sample with replacement from a finite population.

4.0 CONCLUSION

Sampling distribution of a sample statistic is very important in statistical inference. Therefore, understanding this important concept would aid the understanding of hypothesis testing for decision making.

5.0 SUMMARY

A quantity that can be used to summarize the properties of a population is known as parameter, while similar quantity that can be used to summarize the properties of a sample is known as sample statistic. A sampling distribution of a statistic is defined as the distribution of the values of the statistic over all possible repeated samples of same size drawn from a specific population in consideration.

6.0 TUTOR-MARKED ASSIGNMENT

1. Differentiate between a population and a sample
2. Explain the concept of sampling distribution
3. Outline the process of estimation
4. Distinguish between point estimation and interval estimation.
5. Suppose that the averages of all weights of students in a university are normally distributed with mean 42kg and standard deviation 12kg. If groups of 100 students are randomly selected, what are the mean and standard deviation of the corresponding sampling distribution of the mean? What can be said about the shape of this sampling distribution?

7.0 REFERENCES/FURTHER READING

1. Rose, B. (1995). Fundamentals of Biostatistics (4^{th} ed.), Duxbury Press, ITP, USA.

2. Le, C. T. (2003). Introductory Biostatistics. John Wiley and Sons Publishing, Wiley-Interscience.
3. Department of mathematics, University of Lagos (2015). A First Course in Statistics. Lagos: Nile Ventures.

UNIT 2 ESTIMATION

CONTENTS

1.0 INTRODUCTION

Estimation is the statistical procedure of using the value of a sample statistic as an estimate of a population parameter. Researchers and practitioners in health sciences are often interested in population parameters. But, most time they cannot compute a value for the parameter of interest, because population data are not always available. So they get sample data, compute the corresponding statistic of interest for the sample and use it to represent the parameter of interest. In this unit, we shall consider two important aspect of estimation, point estimation and interval estimation.

2.0 OBJECTIVES

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

i. Outline the process of estimation

ii. Distinguish between point estimation and interval estimation

3.0 MAIN CONTENT

3.1 POINT ESTIMATION

Considering the example of a study to investigate the average weight of HIV patients aged 20 to 60 years, we might decide to randomly select a select of 50 HIV patients in the and measure their weights. Suppose the average weight of the 50 patients is, say, 58kg.Then average weight of HIV patients aged 20 to 60 years is the parameter, which a value is needed for, the sample size is 50 and the value of the sample statistic (the average weight of the 50 patients) is 58kg. Then the estimate of the population parameter is 58kg. It can be concluded that average weight of HIV patients aged 20 to 60 years is estimated as 58kg.

In the above procedure, it is noted that just a single value of a corresponding sample statistic (a point) is used as an estimate of the parameter of interest. So, the estimation process is regarded as point estimation. In estimation, the sample statistic is referred to as estimator while its value is the estimate. So, a point estimate is the single numerical value computed from sample data and used as estimate of the population parameter of interest.

3.2 INTERVAL ESTIMATION

Sometimes, there may be need to use a range of values as estimates of the population parameter of interest. In this case, the estimation procedure is defined over an interval and is thus referred to as confidence interval or interval estimation. Let a parameter be denoted by θ and $\hat{\theta}$ the point estimator (the sample statistic whose value is used for estimating θ). Then the $(1-\alpha)$% confidence interval for estimating θ is given by:

$\hat{\theta} \pm SE(\hat{\theta}) \times \delta_{\alpha/2}$;

where $SE(\hat{\theta})$ is the standard error of $\hat{\theta}$, $\delta_{\alpha/2}$ is the critical value of the sampling distribution (usually obtained as tabulated values of the distribution) of $\hat{\theta}$, $(1-\alpha)$ is the probability level or confidence coefficient, specifying the degree of certainty or confidence, that the true value of the parameter θ lies within the given interval and the term ($SE(\hat{\theta}) \times \delta_{\alpha/2}$) is the margin of error. In other words, a confidence interval estimates an interval (*a*, *b*) within which the true value of the parameter will fall given certain confidence level or confidence coefficient; where the end points *a* and *b* are called confidence limits. In the interval (*a*, *b*), *a* is the lower confidence limit and *b* is the upper confidence limit. The confidence level is always given in terms of probability and the most commonly used probability levels are 90%, 95% and 99%. In medical practice and research the acceptable confidence level is 95%.

Example: Consider a pediatrician who estimates the average monthly expenditure on children health care in a state (in ₦M) as ₦40 ± ₦2 using a confidence level of 95%. The pediatrician is actually claiming that he is 95% sure that the actual average monthly expenditure on children health care is in the interval (₦38M, ₦42M), where the lower confidence limit is 40- 2 = 38 and the upper limit is 40+2 = 42. This means that the actual average expenditure on children health care for the state μ lies between ₦38M and ₦42M and can also be written as ₦360 $\leq \mu \leq$ ₦400. Another way of interpreting this is to say that, if repeated samples of the monthly expenditure on children health care for the state are taken, each time computing the average, then it is expected that 95 percent of such sample averages will fall within the interval (₦38M, ₦42M).

4.0 CONCLUSION

The process of estimating population parameter using the corresponding sample statistic has been presented. A distinction and clear explanation of point estimation and interval

estimation has been made. Point estimation is the process whereby a single numerical value is computed from sample data and used as estimate of the population parameter of interest. On the other hand, confidence interval is an estimation procedure where a range of values in an interval is obtained using sample data and used as possible is estimates for population parameter, together with a confidence level that the actual or true value of the population parameter is within the given interval. As pointed out already, a point estimate is needed to obtain the corresponding confidence interval.

5.0 SUMMARY

Estimation is the statistical procedure of using the value of a sample statistic as an estimate of a population parameter. If a single value of the sample statistic is used as an estimate of a population parameter the estimation process is referred to as point estimation.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define estimation
2. Distinguish between point estimation and interval estimation
3. It is known that the average salary of 10 medical professionals in a small town is #100,000. What would be the estimated average salary of all the medical professionals in the small town?
4. A random sample of 400 cars crossing a bridge into a city during rush hour contain an average of 4 people per car with a standard deviation of 1. Find a 95% confidence interval for the average number of people per car on the bridge during rush hour.

5. Assume n = 36, x = 25 and s = 5. Find a 90% confidence interval for mean.

6. Out of 400 randomly selected products in a manufacturing company, 6 are defective. Find a 90% confidence interval for the proportion of all defective items.

7.0 REFERNCES/FURTHER READING

i. Rose, B. (1995). Fundamentals of Biostatistics (4th ed.), Duxbury Press, ITP, USA.

ii. Le, C. T. (2003). Introductory Biostatistics. John Wiley and Sons Publishing, Wiley-Interscience.

iii. Department of mathematics, University of Lagos (2015). A First Course in Statistics. Lagos: Nile Ventures.

UNIT 3 SAMPLING DISTRIBUTIONS OF THE SAMPLE MEAN AND PROPORTION

CONTENTS

6.0 Tutor-Marked Assignment

7.0 References/Further Reading

1.0 INTRODUCTION

The sampling distributions of the sample mean will be treated for both small and large samples. Also, the sampling distribution of sample proportion will be treated here. Most common sampling distributions are the standard normal and t distributions and these will be considered.

2.0 OBJECTIVES

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

i. Identify common distributions used as sampling distributions.

ii. Find the sampling distribution of the sample mean when sample size is large and when sample size is small.

iii. Obtain the sampling distribution of a sample proportion

iv. Obtain the sampling distribution of the difference between two sample means and two sample proportions

3.0 MAIN CONTENT

3.1 COMMON SAMPLING DISTRIBUTIONS

3.1.1 The z distribution and t distribution

The z distribution is the standard normal distribution already covered in previous module. Data on most real life situations follow the normal distribution, which can be standardized into z scores. Therefore, the z distribution is a special type of normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$; usually written as $Z \sim N(0, 1)$. Let X be a normal random variable with mean μ and variance σ^2, then

$Z = Z = \frac{X - \mu}{\sigma}$ is a standard normal random variable.

On the other hand the t distribution, often referred to as the student's t distribution, is a distribution that is indexed by its degrees of freedom and depends on the sample size n. The degrees of freedom is the only parameter of the t distribution.

It is important to note that both the z and t distributions are symmetric distributions. That is, dividing the distribution curve along the center, where x = 0, the area to the right is the same as the area to the left.

Depending on certain conditions, both the z and t distributions are the sampling distribution of the repeated samples of statistics, like the sample mean and proportion.

3.1.2 The F Distribution and Chi Square (χ^2) Distribution

Other sampling distributions frequently encountered in statistical works are the chi square and F distributions. Interestingly, they are both skewed (non-symmetric) distributions and are used in different settings. The chi square distribution has one parameter, which is its degrees of freedom and is used mostly as sampling distribution of squares or sums of squares or variances. On the other hand, the F distribution is a ratio of two chi square distributions and has two parameters, which are the degrees of freedom of the chi square distributions. The F distribution is used as sampling distribution of the ratio of two variances to compare means and variances.

3.2 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n. The sample mean is given by $\frac{1}{n}\sum_{i=1}^{n} X_i$. As mentioned earlier, in Statistics, the sample mean is an unbiased estimator of the population mean.

<u>Case 1</u>: Sampling Distribution of Sample mean ($\bar{X}$) when the population variance is known

Let a random sample be selected from a normal population with known mean μ and variance σ^2. Then whether the sample size is large or small, the following holds for the sample mean $\bar{X}$.

$E(\bar{X}) = \mu$,

$Var(\bar{X}) = \sigma^2/n$,

$SE(\sigma^2) = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$;

The sampling distribution of the sample mean $\bar{X}$ is a normal distribution with mean μ and varianceσ^2/n, where σ is the population standard deviation.

Example 1:

A random sample of size n = 49 patients was selected for a study from a population with mean μ = 35 and standard deviation σ = 5. Find the(i) mean, (ii) variance and (iii) sampling distribution of the sample mean $\bar{X}$**.** (iv) What is the standard error of $\bar{X}$.

Solution:

The mean of $\bar{X}$ is $E(\bar{X}) = 35$,

The variance of $\bar{X}$ is $Var(\bar{X}) = 5/\sqrt{49} = 5/7 = 0.714$,

The sampling distribution of $\bar{X}$ is normal distribution.

The standard error of $\bar{X}$ is $\sqrt{0.714} = 0.845$.

Example 2:

Given that the distribution of a random variable X is normal with parameters $\mu = 10$ and $\sigma^2 = 36$, find (i) $E(\bar{X})$ and (ii) $Var(\bar{X})$ if a sample of size 6 is selected from this population.

Hence, obtain the standard error of $\bar{X}$ and draw conclusion about the sampling distribution of $\bar{X}$.

Solution:

(i) $E(\bar{X}) = \mu = 10$

(ii) $Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{36}{6} = 6$

Furthermore, the standard error of $\bar{X}$ can be deduced as:

$SE(\bar{X}) = \sqrt{Var(\bar{X})} = \sqrt{6} = 2.45$

The sampling distribution of $\bar{X}$ is normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

<u>Case 2</u>: Sampling Distribution of Sample mean ($\bar{X}$) when the population variance is not known and the sample size is large.

If the population variance is not known but the sample size is large *($n \geq 30$)*, the population variance can be replaced with the sample variance (its point estimator) given by $S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(X - \bar{X})^2$. Then the following will hold:

$E(\bar{X}) = \mu$

$Var(\bar{X}) = S^2/n$

$SE(\bar{X}) = \sqrt{S^2/n} = S/\sqrt{n}$;

The sampling distribution of the sample mean $\bar{X}$ is approximately normal distribution with mean μ and variance $Var(\bar{X}) = \frac{S^2}{n}$, where S is the sample standard deviation.

Example 3:

Let the birth weights of babies follow a normal distribution with mean μ = 3kg but unknown variance. Suppose that a random sample of 35 babies is taken from this population and its standard deviation computed as S = 1.8kg. Find the mean, variance, standard error of the sample mean, $\bar{X}$ and its sampling distribution.

Solution:

The mean of $\bar{X}$ = E($\bar{X}$) = 3

The variance of $\bar{X}$ = Var($\bar{X}$) = S^2/n = $1.8^2/35$ = 3.24/35 = 0.0926

Standard error of $\bar{X}$ = $\sqrt{Var(\bar{X})}$ = $\sqrt{0.0926}$ = 0.304

Applying the CLT, the sampling distribution of $\bar{X}$ is normal distribution with mean 3 and standard deviation 0.304.

For the two cases above, the z-score can be computed as $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ for case 1 and $Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ for case 2. You have learnt about z-score as a value of standard normal random variable.

Example 4:

For the problem in Example 3, find the probability that the sample mean will not be greater than 4.

Solution:

From the problem in Example 3, the sampling distribution of $\bar{X}$ is a normal distribution with mean $\mu = 3$ and standard deviation S = 0.304.

Using the fact that $P(\bar{X} \le \bar{x}) = P\left(Z \le \frac{\bar{x} - \mu}{s/\sqrt{n}}\right)$, where Z is a standard score, we have;

$P\left(Z \le \frac{4-3}{0.304}\right) = P(Z \le 3.29) = 0.999$.

<u>Case 3</u>: Sampling Distribution of Sample mean ($\bar{X}$) when the population variance is not known and the sample size is small

If the population variance is not known and the sample size is small *(n < 30)*, the population variance is still replaced with the sample variance given by $S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(X - \bar{X})^2$. Then the following will hold:

E($\bar{X}$) = μ

Var($\bar{X}$) = S^2/n

SE($\bar{X}$) = $\sqrt{S^2/n} = S/\sqrt{n}$;

The sampling distribution of the sample mean $\bar{X}$ is a student's *t* distribution with n-1 degrees of freedom.

For this case, instead of z-score, the t-score would be computedas $t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$, but the t-score approaches the z score as the sample size n increases.

However, it is always good to keep number of samples large enough so that the central limit theorem (CLT) will apply.

Example 5:

A random sample of *n* = 20 infants with dental defect was obtained from a population with mean

= 75. The sample standard deviation was found to be s = 8. Find:

i. E($\bar{X}$), Var($\bar{X}$), SE($\bar{X}$)
ii. The sampling distribution of $\bar{X}$.

Solution:

i. E($\bar{X}$) = 75; Var($\bar{X}$) = $\frac{s^2}{n} = \frac{8^2}{20} = \frac{64}{20}$ = 3.2; SE($\bar{X}$) = $\sqrt{Var(\bar{X})} = \sqrt{3.2}$ = 1.789.

ii. The sampling distribution of $\overline{X}$ is a student's *t* distribution with n-1 = 19 degrees of freedom.

Example 6:

Samples of 25 are drawn from a normal population with mean 85. If the sample standard deviation is 15 find the t-scores (t values) corresponding to the following sample means.

$\overline{X} = 88,$

$\overline{X} = 81$

Solution:

Since the sample size is 25 < 30 and the parent population is normal, the sampling distribution of

$\overline{X}$ is a t distribution with n-1 = 24 degrees of freedom. So $\mu = 85$, s = 15 and

$\frac{s}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 15/5 = 3.$

For $\overline{X} = 88,\ t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{88-85}{3} = \frac{3}{3} = 1.00.$

For $\overline{X} = 81,\ t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{81-85}{3} = \frac{-4}{3} = -1.33.$

The Central Limit Theorem:

As shown above, if the underlying distribution of the sample is normal with mean μ and variance σ^2, then the sampling distribution of the sample mean $\overline{X}$ would be normal; i.e. $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$

The challenge is that many of the distributions encountered in practice are not normal. In such situation, an important theorem in Statistics known as the Central Limit Theorem (CLT) always come to play. The theorem states that when a random sample of size n is drawn from any population with mean μ and variance σ^2, irrespective of the underlying distribution, $\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$, provided n is sufficiently large.

Example 7:

A stethoscope manufacturer claims that its stethoscopes will last an average of 15 years with a standard deviation of 7 years. A random survey of 36 stethoscopes was collected. Find the probability that the sample mean $\overline{X}$ of the 36 stethoscopes is less than 12 years.

Solution:

Applying CLT (since the parent distribution is not specified but n is large enough), the sampling distribution of $\overline{X}$ is approximately normal. So, the mean $\mu = 15$, n = 50, $\sigma = 7$, and the z score for the mean is $z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{12-15}{7/\sqrt{36}} = \frac{-3}{1.167} = -2.57$.

$\therefore P(\overline{X} < 12) = P(Z \leq -2.57) = 0.0051.$

3.3 SAMPLING DISTRIBUTION OF SAMPLE PROPORTION

Let X_1, X_2, …, X_n be a random sample of size *n*, some of which have an attribute of interest. Let *m* out of the *n* sample members have the attribute. Then the sample proportion is defined as p = $\frac{m}{n}$. The sample proportion p is an unbiased estimator of the corresponding population proportion P.Accordingly,

The mean of p, E(p) = P.

The variance of p is Var(p) = $\frac{PQ}{n}$

The standard error of p is given by se(p) = $\sqrt{\frac{PQ}{n}}$, where P is the population proportion, Q = 1-P and n the sample size.

N/B: In a situation where the population proportion is not available, it can be estimated by the sample proportion p, since the sample proportion is an unbiased estimate of the population proportion.

The sampling distribution of population proportion is normal with mean P and variance $\frac{PQ}{n}$.

Example 8:

Suppose in a random sample of 100 people in a local district, it was found that 60 of them test positive to the corona virus. Estimate the proportion P of people in the local district that would test positive to the corona virus together with its standard error.

Solution:

Estimate of P, the population proportion of people in the district with corona virus, is the sample proportion p.

$p = 60/100 = 0.6.$

$$\text{Var}(p) = \frac{pq}{n} = \frac{0.6 \times 0.4}{100} = \frac{0.24}{100} = 0.0024,$$

$$\therefore \text{SE}(p) = \sqrt{0.0024} = 0.049.$$

3.4 SAMPLING DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO SAMPLE MEANS AND PROPORTIONS

3.4.1 Difference between Two Sample Means

If two independent random samples of sizes n_1 and n_2 are drawn from two normally distributed populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively, the sampling distribution of the difference between the two sample means $(\overline{X}_1 - \overline{X}_2)$ is a

normal distribution with mean $\mu_{\bar{X}_A-\bar{X}_B} = \mu_A - \mu_B$ and variance $\sigma^2{}_{\bar{X}_A-\bar{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}$.

From the variance of the difference stated above, one can deduce that the standard error of the difference between two sample means is $\sigma_{\bar{X}_A-\bar{X}_B} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$.

Example 9:

Supposing there are two groups, A and B, of women visiting an ante-natal clinic. The mean height of women in group A is 32 while the mean height of women in group B is 22. The variances of the two groups are 60 and 70, respectively and the heights of women in both groups are normally distributed. A medical researcher randomly sample 10 womenfrom group A and 14 from group B.

Determine the sampling distribution of the difference between the means of sample A and B

What is the probability that the mean height of the 10 women from group A will exceed the mean height of the 14 women from group B by 5 or more?

Solution:

To determine the sampling distribution of the difference between the two means, use the formulas above,

The mean of the difference is:

$\mu_{\bar{X}_A-\bar{X}_B} = \mu_1 - \mu_2 = 32 - 22 = 10$,

The standard error is:

$$\sigma_{\bar{X}_A-\bar{X}_B}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}=\sqrt{\frac{60}{10}+\frac{70}{14}}=3.317.$$

So, the sampling distribution of the difference is normal distribution with mean 10 and standard deviation 3.317.

From the above calculations, we know that the difference in population means is 10. Therefore, there is a high probability that the mean height of the women from group A will exceed that of group B by 5 or more? We are interested in finding the exact probability? Using standard normal probability table, we have

$$P\left[(\bar{X}_A-\bar{X}_B)\geq 5\right]=1-P\left[(\bar{X}_A-\bar{X}_B)<5\right]=1-P\left(Z<\frac{5-10}{3.317}\right),$$

$$=1-P(Z<-1.51)=1-0.0655=0.9345.$$

If in the above formulas, the two groups have same population variance, say σ^2, and same sample size *n*. Then the sampling distribution of the difference between the two sample means would still be normal with mean $\mu_{\bar{X}_A-\bar{X}_B}=\mu_A-\mu_B$ and variance

$$\sigma^2{}_{\bar{X}_A-\bar{X}_B}=\frac{\sigma^2}{n}+\frac{\sigma^2}{n}=\frac{2\sigma^2}{n}.$$

Example 10:

The mean weight of 15-year-old boys (in kg) is 175 and the variance is 64. For girls, the mean is 165 and the variance is 64. If eight boys and eight girls were sampled, what is the probability that the mean height of the sample of girls would be higher than the mean height of the sample of boys?

Solution:

Let G be sample of girls and B sample of boys. Like before, the problem can be solved by first finding the sampling distribution of the difference between means ($\overline{G} - \overline{B}$). The mean of the sampling distribution is 165 - 175 = -10. The standard deviation of the distribution is:

$$\sigma_{\overline{G}-\overline{B}} = \sqrt{\frac{2\sigma^2}{n}} = \sqrt{\frac{2\times 64}{8}} = \sqrt{16} = 4.$$

Therefore, the sampling distribution of the difference between means is normal with mean -10 and standard deviation 4.

So, $P(\overline{G} > \overline{B}) = P\left((\overline{G} - \overline{B}) > 0\right) = P\left(Z > {}^{10}\!/_{4}\right),$

$$= 1 - P(Z \le 2.5) = 1 - 0.9938 = 0.0062.$$

Notice that if the population variance is not known in the above formulas, it can be estimated by the sample variance. However, if the sample size is less than 30 in this case the sampling distribution would be a t distribution.

3.4.1 Difference between Two Proportions

Similarly, the sampling distribution of the difference between two sample proportions is approximately normal with mean $\mu_{p_1-p_2} = P_1 - P_2$, difference in proportions of population 1 and population 2, and standard error is given as $\sigma_{p_1-p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$; $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Note: If the population proportions P are not known, the sample proportions p will be used as good replacement for the population proportions.

Example 11

A random sample of size 200 was selected from male in a state capital and 25 were found to test positive to HIV. A second sample of size 300 was selected from among female and 30 were found to test positive. Compute the

difference between the population proportions

standard error of the differences between the population proportions.

Solution

From the question, $n_1 = 200$, $X_1 = 25$, $n_2 = 300$, and $X_2 = 30$. Therefore,

$\bar{P}_1 = \frac{25}{200} = 0.125$; $\bar{P}_2 = \frac{30}{300} = 0.10$

(i) Difference between the two population proportions is $\bar{P}_1 - \bar{P}_2 = 0.125 - 0.100 = 0.025$

(ii) Standard error of the differences between the two population proportions is

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = \sqrt{\frac{(0.125)(0.875)}{200} + \frac{(0.1)(0.9)}{300}} = \sqrt{0.000547 + 0.0003} = 0.029$$

4.0 CONCLUSION

In this unit, we have considered the sampling distribution of the sample mean and proportion. Detailed explanations have been given for obtaining the sampling distributions of the sample mean for large and small samples as well as sample proportion. The sampling distributions of the difference between two sample means and sample proportions have been considered too.

5.0 SUMMARY

The summary of this unit are as follows. If the parent population is normal with known variance, the sampling distribution of the sample mean and proportion is a normal

distribution. If, however, the variance is not known and the sample size is small the sampling distribution would be a t distribution. If the parent population is not known, the central limit theorem could be applied by keeping the sample size large.

6.0 TUTOR-MARKED ASSIGNMENT

1. List some common sampling distributions

2. 8. A recent study has found that the mean number of hours per week that people work in Nigeria is 50 with a standard deviation of 9 hours.

a.) If groups of 32 people are randomly selected, what are the parameters of the distribution of sample means? What can you conclude about the shape of this distribution?

b.) If groups of 25 people are randomly selected, what are the parameters of the distribution of sample means? What can you conclude about the shape of this distribution?

3. Consider the approximately normal population of weights of female students at a college. Assume that the individual weights have a mean of 61kg and a standard deviation of 10kg. A random sample of 25 weights is obtained.

i. Find the mean of this sampling distribution
ii. Find the standard error of the mean
iii. Find the shape of this sampling distribution
iv. Find $P(x > 65)$
v. Find $P(x < 60)$

4. Population 1 has a mean of 20 and a variance of 100. Population 2 has a mean of 15 and a variance of 64. You sample 20 scores from Pop 1 and 16 scores from Pop 2. What is the mean of the sampling distribution of the difference between means (Pop 1 - Pop 2)?

5. Given that two independent and normally distributed populations with means 54 and 58; variances 36 and 64 respectively; and also given that two independent samples of sizes 10 and 8 were chosen from the two populations. Find (i) the sampling distribution of the difference between the means (ii) the standard error of the differences between the two sample means.

6. A political poll conducted in a local government area reveals that 65 out of 100 randomly selected people surveyed voted for a candidate. Estimate the proportions of people who voted for the candidate with its standard error.

7.0 REFERNCES/FURTHER READING

1. Rose, B. (1995). Fundamentals of Biostatistics (4^{th} ed.), Duxbury Press, ITP, USA.
2. Le, C. T. (2003). Introductory Biostatistics. John Wiley and Sons Publishing, Wiley-. Interscience.
3. Department of mathematics, University of Lagos (2015). A First Course in Statistics. Lagos: Nile Ventures.

UNIT 4 CONFIDENCE INTERVALS FOR POPULATION MEAN AND PROPORTION

CONTENTS

1.0 INTRODUCTION

Once the sampling distribution of a sample statistic is ascertained a confidence interval can be made about the corresponding population parameter. Recall that the $(1-\alpha)$% confidence interval for estimating a population parameter θ is given by:

$$\hat{\theta} \pm SE(\hat{\theta}) \times \delta_{\alpha/2};$$

Where $\hat{\theta}$ is the point estimator (statistic used in producing the estimate) of θ, $SE(\hat{\theta})$ is the standard error of $\hat{\theta}$, $\delta_{\alpha/2}$ is the critical value of the sampling distribution of $\hat{\theta}$(you will learn more about critical values later), $(1-\alpha)$is the probability that the estimated interval actually contains the true value of the parameter.

2.0 OBJECTIVES

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

i. Estimate confidence Interval for the Mean of a Population with Known Variance and/or Large Sample Size

ii Obtain confidence Interval for the Mean of a Population With Unknown Variance and Small Sample Size

iii Find confidence Interval for Population Proportion

3.0 MAIN CONTENT

3.1 CONFIDENCE INTERVAL FOR THE MEAN OF A POPULATION WITH KNOWN VARIANCE AND/OR LARGE SAMPLE SIZE

The sampling distribution of the sample mean $\overline{X}$ can lead to obtaining a sampling distribution of the statistics, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$, an important statistic used in deriving confidence interval and generally used in making inference about the population mean μ .Notice that the denominator is the standard error of $\overline{X}$ .The sampling distribution of the statistic $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ and/or $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ is normal when σ is known and/or sample size large. However, the confidence interval for the population mean would be stated here without derivation.

Let μ be the mean of a normal population and $\overline{X}$ the point estimator of μ. Then the $(1-\alpha)$% confidence interval for estimating μ is given by:

$$\overline{X} \pm SE(\overline{X}) \times Z_{\alpha/2};$$

where $SE(\overline{X})$ is the standard error of $\overline{X}$, $Z_{\alpha/2}$ is the critical value of the standard normal distribution, being the sampling distribution of $\overline{X}$ .For the medical field, α is always given at 0.05, so that $(1-\alpha)$% is 95%.

If the population variance σ^2 is known, then $SE(\overline{X}) = \frac{\sigma}{\sqrt{n}}$. If the population variance σ^2 is not known and sample size is large (greater than 30), then σ can be replaced by the

sample standard deviation S, since S is an unbiased estimator of σ, in which case $\text{SE}(\overline{\text{X}}) = \frac{\text{S}}{\sqrt{\text{n}}}$, and the sampling distribution of $\overline{\text{X}}$ is approximately standard normal.

Example 1:

A random sample of 400 cars crossing a bridge into a city during rush hour contain an average of 4 people per car with a standard deviation of 1. Find a 95% confidence interval for the average number of people per car on the bridge during rush hour.

Solution:

$\text{n} = 400,\ \overline{\text{X}} = 4,\ \text{S} = 1$

$$\therefore \text{SE}(\overline{\text{X}}) = \frac{S}{\sqrt{n}} = \frac{1}{\sqrt{400}} = \frac{1}{20} = 0.05,$$

$$95\% \Rightarrow \alpha = 0.05,$$

$$\therefore Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96,$$

Therefore, the 95% confidence interval for μ is

$$\overline{\text{X}} \pm \text{SE}(\overline{\text{X}}) \times Z_{\alpha/2} = 4 \pm 0.05 \times 1.96 = 4 \pm 0.098,$$

$$= (3.902, 4.098).$$

Example 2:

The average monthly profit of 64 hospitals operating in Lagos State is ₦18million with population standard deviation of ₦4 million. Construct a 95% confidence interval for the true monthly average profit of the hospitals.

Solution:

Here, we are given sample size $\bar{X}$ = 18, sample size (n) = 64, population standard deviation (σ) = 4 and α = 5%. Therefore, $Z_{\frac{\alpha}{2}} = Z_{0.025}$ = 1.96 from the normal table.

The confidence interval for $\mu = \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 18 \pm 1.96\left(\frac{4}{8}\right)$ = (17.02, 18.98)

3.2 CONFIDENCE INTERVAL FOR THE MEAN OF A POPULATION WITH UNKNOWN VARIANCE AND SMALL SAMPLE SIZE

If the population variance σ^2 is not known, it could be estimated by the sample variance S^2 and, as stated earlier, for large sample the sampling distribution of the statistic $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ is normal. However if the population variance σ^2 is not known and sample size is small (less than or equal to 30), then σ can be replaced by the sample standard deviation S, since S is an unbiased estimator of σ, but the sampling distribution of the statistic $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ would no longer be normal, since the sampling distribution of $\bar{X}$ is not normal. The sampling distribution has been shown to follow a t distribution with n-1 degrees of freedom.

Therefore,$(1-\alpha)$% confidence interval of the population mean μ is given by:

$$\bar{X} \pm SE(\bar{X}) \times t_{\alpha/2,(n-1)};$$

where $SE(\bar{X}) = S/\sqrt{n}$ is the standard error of $\bar{X}$, $t_{\alpha/2,(n-1)}$ is the critical value of the t distribution (tabulated values on the t distribution table).

Example 3:

A sample of 16 private hospitals in Lagos state shows a mean cost of NHIS services of ₦25000 with a standard deviation of ₦4500. Find a 95% confidence interval the mean cost of NHIS services of private hospitals in the state.

Solution:

The given sample statistics are n = 16, $\overline{X}$ = 25000, s = 4500, α = 1-0.95 = 0.05

$\therefore \text{SE}(\overline{X}) = S/\sqrt{n} = 4500/\sqrt{16} = 4500/4 = 1125,$

And, $t_{\frac{0.05}{2};(16-1)} = t_{0.025;15} = 2.131.$

Let the mean cost of NHIS services of private hospitals in the Lagos state be μ. Then the 95% confidence interval for μ is:

$$\overline{X} \pm \text{SE}(\overline{X}) \times t_{\alpha/2,(n-1)} = 25000 \pm (1125)(2.131) = 25000 \pm 2397.375 = (22602.625, 27397.375).$$

.

3.3 CONFIDENCE INTERVAL FOR POPULATION PROPORTION

The $(1-\alpha)$% confidence interval of the population proportion P is given by:

$$p \pm Z_{\alpha/2}\text{SE}(p);$$

Where p is the sample proportion, $Z_{\alpha/2}$ is the critical value of the standard normal distribution and SE(p) is the standard error of p.

Example 4:

A survey of 100 randomly selected employees of an organization show that 35 drive themselves to work. Find a 90% confidence interval for the proportion of employees P that drive themselves to work in the organization.

Solution:

Given: n = 100, x = 35, α = 1-0.90 = 0.10. Therefore, the sample proportion p = 35/100 – 0.35.

$$\text{SE}(p) = \sqrt{p(1-p)/n} = \sqrt{0.35 \times 0.65/100} = \sqrt{0.002275} = 0.047697; Z_{\alpha/2} = Z_{0.1/2} = Z_{0.05} = 1.645.$$

Therefore, the 90% confidence interval for P is:

$$p \pm Z_{\alpha/2}\text{SE}(p) = 0.35 \pm (1.645)(0047697) = 0.35 \pm 0.078462 = (0.2715, 0.4285).$$

CONCLUSION

In this unit, we have considered the confidence interval for a population mean and proportion. Many illustrative examples have been provided for better understanding.

5.0 SUMMARY

Once the sampling distribution of a sample statistic is ascertained a confidence interval can be made about the corresponding population parameter. Specifically, sampling distributions of the sample mean and sample proportion have been established and used in confidence interval estimation of population mean and proportion.

6.0 TUTOR-MARKED ASSIGNMENT

1. A random sample of 400 cars crossing a bridge into a city during rush hour contain an average of 4 people per car with a standard deviation of 1. Find a 95% confidence interval for the average number of people per car on the bridge during rush hour.

2. Assume n = 36, x = 25 and s = 5. Find a 90% confidence interval for mean.

3. Out of 400 randomly selected products in a manufacturing company, 6 are defective. Find a 90% confidence interval for the proportion of all defective items.

7.0 REFERNCES/FURTHER READING

i. Rose, B. (1995). Fundamentals of Biostatistics (4^{th} ed.), Duxbury Press, ITP, USA.

ii. Le, C. T. (2003). Introductory Biostatistics. John Wiley and Sons Publishing, Wiley-Interscience.

iii. Department of mathematics, University of Lagos (2015). A First Course in Statistics.

Lagos: Nile Ventures.

MODULE 2 TEST OF HYPOTHESIS

Unit 1 Concepts in Testing Hypothesis
Unit 2 Test for Mean and Proportion of One and Two Samples
Unit 3 One Way ANOVA and Chi Square Test
Unit 4 Correlation and regression Analysis

UNIT 1 CONCEPTS IN HYPOTHESIS TESTING

CONTENT

1.0 INTRODUCTION

In an attempt to make valid decisions about populations based on sample information, it is useful to make assumptions or guesses about the populations involved. Such assumptions, which may or may not be true, are called statistical hypotheses. Statistical hypotheses are statements about the probability distributions of the populations and these statements need to be validated. The statistical procedures for validating the truth or falsity of a hypothesis or to determine whether observed samples differ significantly from expected results are called tests of hypotheses or tests of significance. In this unit, we define basic terminologies health managers and researchers can encounter while testing hypothesis and also discuss the steps involved in testing various types of hypotheses so as to arrive at valid decisions.

2.0 OBJECTIVES

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

i. Define some basic concepts relating to hypothesis testing
ii. Formulate null and alternative hypotheses for testing.
iii. Define errors committed in a statistical experiment.
iv. Enumerate steps in hypothesis testing.

3.0 MAIN CONTENTS

3.1 DEFINITIONS OF BASIC TERMS

The Null and Alternative Hypotheses: In hypothesis testing, the null hypothesis denoted H_0, is formulated that there is no difference between the procedures (that is, any observed differences are merely due to fluctuations in sampling from the same population). It is a hypothesis of primary interest, in which the intention is to reject it at the end of the test, assuming convincingly that it is wrong. This is why it is always stated in negation.

The alternative or research hypothesis denoted by H_1 is a statement that is to be accepted, assuming there is convincing sample evidence that it is true. It is always stated in affirmative.

Types of Errors: Sometimes a hypothesis can be wrongly accepted of wrongly rejected. This would lead to errors in decision making. Two types of errors are possible: type I and Type II errors. A type I is committed when a true null hypothesis is rejected and Type II error is error committed when a false null hypothesis is accepted. When no error is committed, a correct decision is taken. This is summarized in the table below.

Table 8.1: Decision Table

H_0	Accept	Reject
True	Correct decision	Type I error
False	Type II error	Correct decision

Significance Level: Significance level, often denoted by α, is the highest probability of rejecting the null hypothesis. It is the probability associated with type I error and is a measure of the error level one can tolerate in any test of hypothesis. Since it is not always possible to avoid error completely, the best one can do is to control the risk or probability of making such errors. On the other hand, the probability associated with type II error is often denoted by β and it is just the compliment of α.Commonly used significance levels are 0.1, 0.05, 0.025, 0.01; but for the medical profession, 0.05 is advisable.

Test Statistic: A test statistic is a statistic (a random variable) whose value (computed from sample data) is used to take decision whether or not to reject the null hypothesis. This value is always referred to as calculated value or test value. You can view a test statistic as a formula that helps to calculate this test value from sample data.

Critical value: It is the value obtained from the underlying distribution for comparing with the calculated value from the purpose of making decision whether or not to reject H_0. It is also referred to as tabulated value and can easily be obtained from standard statistical tables at a specified significance level.

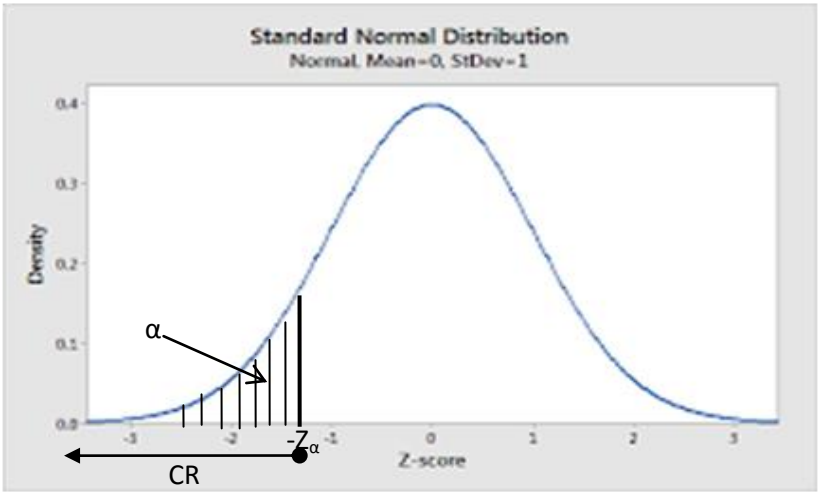
Critical Region: A critical region is an interval or set of values of the random variable (based on its sampling distribution) within which a test statistic would lead to rejecting the null hypothesis. Critical region is also known as rejection region and the compliment is acceptance region.

Direction of a Test: The direction, or sidedness, of a test refers to the side of the distribution where the critical region lies. The direction of a test is determined from the alternative hypothesis. In most of the tests, we will consider the null hypothesis to be an assertion that a population parameter, say μ, has a specific value, say θ, or stated in negative way, it is not different significantly from θ. Mathematically, H_0: $\mu = \theta$. The alternative hypothesis will be one of the following assertions:

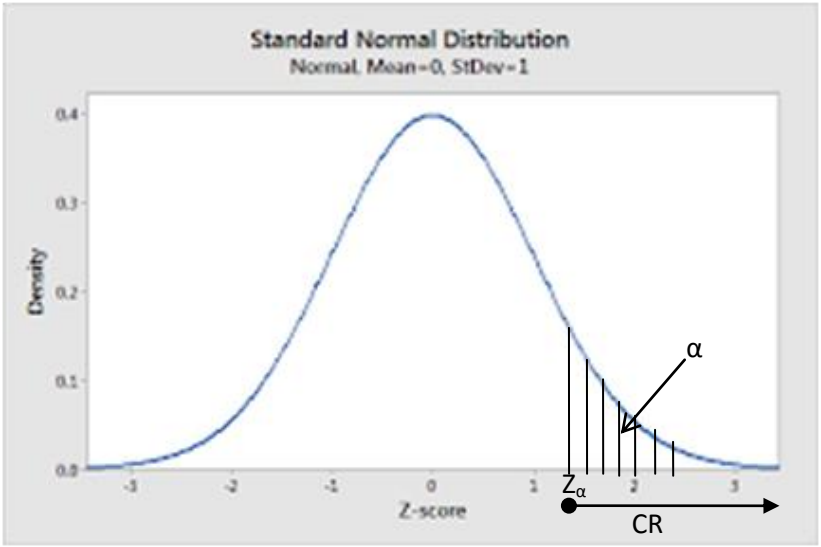
i. The parameter is less than the stated value (H_1: $\mu<\theta$); critical region is on the left side of the distribution, leading to a left-tailed test or left-sided test. This is a one-tailed test.
ii. The parameter is greater than the stated value (H_1: $\mu>\theta$); critical region is on the right side of the distribution, leading to a right-tailed test or right-sided test. This is a one-tailed test.
iii. The parameter is not equal to (greater than or less than) the stated value (H_1: $\mu \neq \theta$); critical region is on both sides of the distribution, leading to a two-tailed test or two-sided test.

The three cases are given as illustration in the figure below. Consider a standard normal distribution as shown below, the x-axis bears values of the random variable Z having the

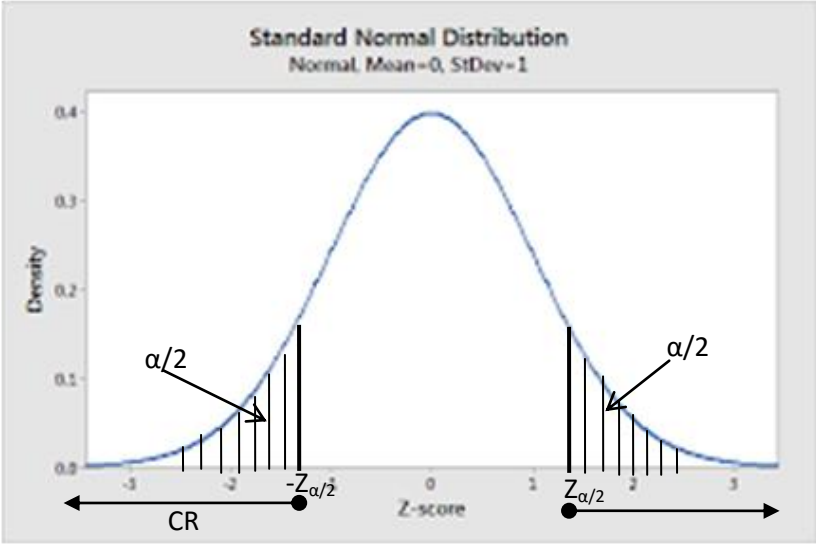
distribution, α is the significance level, Z_α is the critical value (Z value at any given α) and CR is the critical region.



Case 1: The critical value and critical region lie only on one side (left side or left tail) of the distribution. It will lead to a one-sided or one-tailed (left-tailed) test.



Case 2: The critical value and critical region lie only on one side (right side or right tail) of the distribution. It will lead to a one-sided or one-tailed (right-tailed) test.



Case 3: The critical value and critical region lie on both sides (right side or left tails) of the distribution. It will lead to a two-sided or two-tailed test. Notice that α is divided by two, because it is the same probability you have to take a wrong decision.

3.2 TESTING A HYPOTHESIS

3.2.1 Power of Test (1 – β): This is the probability of rejecting H_0 given that H_1 is correct. It is however the complement of the probability of committing type II error. That is the probability of correctly rejecting the null hypothesis.

3.2.1 Simple Hypothesis and Composite Hypothesis: If the values population parameters, such as μ, P or σ^2, are provided the hypothesis being tested is known as simple hypothesis.

Hypothesis being tested is said to be a composite hypothesis, if the population parameters are not provided. For example, $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ where μ_1 and μ_2 are not provided.

3.2.3 Steps in Hypothesis Testing

i. Read or study carefully the available statements and information provided. Based on them, state the null and alternative hypotheses.

ii. Decide on the most appropriate statistic to be used. This is determined by the nature of empirical data available, whether the population parameters are provided and the sampling distribution of the corresponding sample statistic. For example, for a univariate data where the population variance or standard deviation is known, z distribution of the form, $Z_{cal} = \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}}$ will be applied. On the other hand, if either variance or standard deviation is unknown and the sample size is less than 30, the most appropriate statistic should be t–distribution of the form $t_{cal} = \frac{(\bar{x} - \mu)}{s/\sqrt{n}}$ with the corresponding hypothesized t-value read at (n-1) degree of freedom.

iii. Compute the test value, using the test statistic. This is called calculate Value

iv. Determine the critical or rejection region. This is done based on the alternative hypothesis.

v. Using the statistical table, obtain the critical value or tabulated value.

vi. Compare the calculated statistic and the tabulated value.

vii. Take a decision. If the calculated value falls within the critical or rejection region reject the null hypothesis, otherwise, do not reject it.

Important Things to Note

i. The null hypothesis is always expressing equality, signifying no significant difference.

ii. A hypothesis is usually stated using population parameter.

iii. Sample statistic is used in testing a hypothesis, ie., using sample data

iv. Result from testing a hypothesis is used to draw conclusion on the entire population.

4.0 CONCLUSION

In this module, we discussed in details, various concepts and procedures for testing hypothesis. Several examples were used to demonstrate the procedures to allow easy understanding for health practitioners and researchers.

5.0 SUMMARY

Different terms and concepts used in hypothesis testing have been defined and explained. Step by step procedure for testing a hypothesis has also been presented.

6.0 TUTOR-MARKED ASSIGNMENT

1. In testing hypothesis, null hypothesis is accepted when the calculated statistic
   (a) falls into the critical region.
   (b) falls outside the critical region.
   (c) is greater than the table value.
   (d) is equal to the table value.

2. When the equality of more than two population means are being tested, the most appropriate statistical tool to be used is …….
   a. Analysis of variance.
   b. Standard Normal distribution.
   c. χ^2-test for normality.
   d. χ^2-test for dependency.

3. Which of these factors determines the sidedness of the test to be applied, when testing hypothesis?
   a) Level of significance.
   b) Null hypothesis.
   c) Alternative hypothesis
   d) Type I and Type II error

4. The probability of rejecting H_0 given the H_1 is correct is known as . . . …….

5. Any statement in research which is subject to testing in order to decide whether to accept or reject it is known as . . . ……

6. Another name for rejection region is . . . ……..

7. Briefly explain these statistical concepts

(i) Null hypothesis (ii) Alternative hypothesis (iii) Type I error (iv) Type II error (v) Level of significance (vi) Test Statistic (vii) Critical region

7.0 REFERENCES/FURTHER READING

i. Mojekwu J. N. (2012). Business Statistics with Solved Examples, Easy Print Publication, Lagos.

ii. Onyeka-Ubaka, J. N. (2013). *Multi-Level Statistics: An Academic Companion for Inter-Disciplinary Professional Competence*, Royal Choice Multi-Media, Lagos.

iii. Spiegel, A. (2011). Statistics. New York, Schum Series.

iv. Rose, B. (1995). Fundamentals of Biostatistics (4^{th} ed.), Duxbury Press, ITP, USA.

v. Le, C. T. (2003). Introductory Biostatistics. John Wiley and Sons Publishing, Wiley-Interscience. Department of mathematics, University of Lagos (2015). A First Course in Statistics. Lagos: Nile Ventures.

UNIT 2 TEST FOR MEAN AND PROPORTION OF ONE AND TWO SAMPLES

CONTENT

1.0 INTRODUCTION

In this unit, we shall be presenting specific hypothesis testing involving the means and proportion of one sample or two samples. The typical null hypothesis here is that the parameter is a fixed value. Depending on the problem at hand, this null hypothesis can be tested against any of the three alternatives. Let us consider an example of a manager's claim that the average monthly profit is ₦40 million; that is, the null hypothesis is that the average monthly profit is equal to₦40 million, i.e. $H_0: \mu = 40$

The three possible alternatives are:

i. The average monthly profit is more than ₦40 million, i.e. $H_1: \mu > 40$
ii. The average monthly profit is less than ₦40 million ,i.e.$H_1: \mu < 40$
iii. The average monthly profit is not ₦40 million, i.e. $H_1: \mu \neq 40$

In a one sample test, for any null hypothesis, $H_0 : \mu = \mu_0$, the table below provides different critical regions for different alternative hypotheses and corresponding decision rules. Let τ_{cal} be the calculated value and τ_{crit} be the critical value.

S/N	Alternative Hypothesis	Critical Region	Decision Rule
1	$H_1 : \mu > \mu_0$	$\tau_{cal} > \tau_{crit}$	If calculated value is greater than tabulated value, reject H_0
2	$H_1 : \mu < \mu_0$	$\tau_{cal} < -\tau_{crit}$	If calculated value is less than tabulated value, reject H_0
3	$H_1 : \mu \neq \mu_0$	$\tau_{cal} > \tau_{crit}$ or $\tau_{cal} < -\tau_{crit}$	If calculated value greater than tabulated value or if calculated value is less than minus tabulated value, reject H_0; ie, reject H_0 if $\lvert\tau_{cal}\rvert > \tau_{crit}$.

It is also important to state that for any one sample test concerning a population parameter, the test statistic is given by: (sample statistic – mean of sample statistic) divided by standard error of sample statistic

2.0 OBJECTIVES

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

i. Test hypothesis concerning the mean of a population with known variance.
ii. Test hypothesis concerning the mean of a population with unknown variance but sample size is large.
iii. Test hypothesis concerning the mean of a population with unknown variance and sample size is small.
iv. Test hypothesis concerning the mean of a population

3.0 MAIN CONTENT

3.1 ONE SAMPLE TEST FOR THE MEAN

3.1.1 One Sample Test for Mean with Known Population Variance and Large Sample

As noted earlier, this test involves only one sample and is a test concerning the mean of the underlying populationμ. Recall that when the population variance σ^2 is known, the sampling distribution of the sample mean $\overline{X}$ is a z distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ (this explanation will be used to obtain test statistic for other test to be considered). Therefore, the test statistic for test this test is:

$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$; where μ_0 is the given value of μ.

However, if the population variance $\sigma^2$is not known but the sample size is large, the test statistic is: $Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$; where s is the estimate of σ.

Example 1

A random sample of size 16 selected from a normal population with variance 16 gives a sample mean of $\bar{X}$ = 23. Test the hypothesis H_0: μ = 22 against the alternative H_1: μ ≠ 22 at 5% level of significance.

Solution:

Step 1**:** H_0: μ = 22; H_1: μ ≠ 22

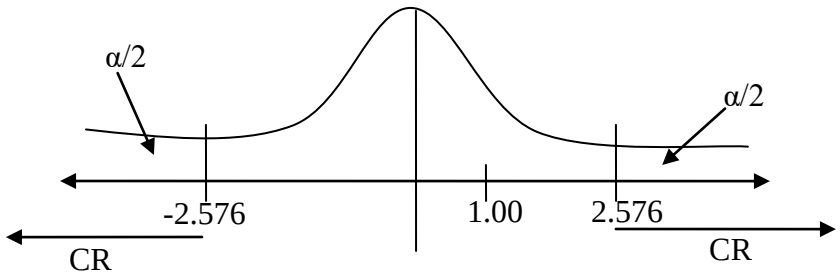
α = 0.05; σ = √16 = 4; μ_0 = 22; $\bar{X}$ = 23; n = 16

Step 2**:** Since σ is known, the test statistic is $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$.

Step 3**:** The test value is:

$$Z_{cal} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{23-22}{4/\sqrt{16}} = 1.00,$$

Step 4: The test is a two-tailed test and hence the critical region as well as the critical value lie on both sides of the distribution. Therefore, α will be split into two, i.e., α/2 = 0.025. So, critical value is $Z_{crit} = Z_{\alpha/2} = Z_{0.01/2} = Z_{0.005} = 2.576$. This is a Z score and will be 2.576 on the right of the distribution and –2.576 on the left (since α was divided by 2).



As can be seen from the figure above, the calculated value, 1.00, is not within any of the critical regions. This statement is equivalent to saying that the calculated value (1.00) is not greater than 2.576 or less than -2.576. Therefore, the null hypothesis cannot be rejected.

Example 2:

Use the given sample data to test the null and alternate hypothesis at a significance level of $\alpha = 0.05$ for the following: H_0: $\mu = 215$; H_1: $\mu < 215$, $n = 49$, $\bar{X} = 200$, $s = 70$.

Solution:

H_0: $\mu = 215$; H_1: $\mu < 215$, $n = 49$, $\bar{X} = 200$, $s = 70$, $\alpha = 0.05$, $\mu_0 = 215$.

Test statistic: In this case, the population standard deviation σ is not known but the sample size

n is large. So, the test statistic is $Z = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$.

Therefore, the test value is: $Z_{cal} = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}} = \dfrac{200-215}{70/\sqrt{49}} = \dfrac{-15}{70/7} = \dfrac{-15}{10} = -1.5$.

Using table, we obtain the critical value Z_α as 1.645.

The critical region is on the left, so decision rule is to reject H_0 if $Z_{cal} < -Z_\alpha$.

Decision: Since $Z_{cal} = -1.5$is not less than $-Z_\alpha = -1.645$, we cannot reject H_0.

Example 3:

It is claimed that the typical Nigeria adult reads an average of 15 books per year. A random sample of 100 adults shows that they read an average of 18 books last year with a standard deviation of 6 books. Test if the stated mean 15 is too low at $\alpha = 0.05$ level of significance.

Solution:

H_0: μ = 15, H_1: μ >15, $\bar{X}$ =18, n = 100, S = 6, α = 0.05.

The test statistic and value are: $Z_{cal} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{18-15}{6/\sqrt{100}} = \frac{3}{6/10} = \frac{3}{0.6} = 5$.

The area of the critical region is 0.05 with a critical value $Z_{0.05}$ = 1.645.

Decision: Since $Z_{cal} = 5 > Z_\alpha = 1.645$, we reject H_0. This means that the stated mean of 15 is too low.

Example 4:

A salesman selected a random sample of size 9 of his daily sales and found the mean to be 6.0. If it is known that themean daily sales of this agent is 5.0 with variance 25. Test the hypothesis that the true population mean daily sale of the agent is greater than 5.0 at α = 0.05.

Solution:

$H_0: \mu = 5.0$against $H_1: \mu > 5.0$, $\bar{x}$ = 0.6, α = 0.05.

Here, although the sample size n = 9 is less than 30, we apply a z test, since population variance σ^2 is given as 25. Therefore,

$$Z = \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} = \frac{(6.0 - 5.0)}{5/\sqrt{9}} = \frac{3}{5} = 0.6$$

Now, the corresponding table value can be obtained using the given level of significance 5%, i.e, $Z_{0.05} = 1.645$ since it is an upper tailed test.

This suggests that the null hypothesis should be accepted as the calculated value is not greater than te critical value.

3.1.2 One Sample Test for Mean with Unknown Population Variance and Small Sample

In this section, we consider data set with a sample size of 30 or less. We cannot use a normal distribution, but rather we must work with an appropriate t-distribution with n – 1 degrees of freedom, provided that the underlying population is approximately normal. The t-values instead of Z values are used. The test statistic is given as

$t_{cal} = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$, S = sample standard deviation.

The test procedure is the popular one sample t test.

Example 5

An administrator claims that the average daily allowances (in N'000) paid to his workers is 6. An accountant attached to the company in an attempt to verify this claim, took a random sample of daily allowances for 10 days and recorded them as 5, 10, 8, 5, 5, 6, 8, 6, 5, 8. Test at 5% level of significance the validity of the administrator's claim.

Solution:

The hypotheses are; $H_0: \mu = 6$ against $H_1: \mu \neq 6$

n = 10, $\alpha = 0.05,\ \sigma\ is\ unknown.$ Hence, this calls for the use of t-distribution. We obtain sample mean and sample standard deviation so as to use the statistic as follows:

X	X^2
5	25
10	100
8	64
5	25
5	25
6	36

$$t_{cal} = \frac{\sqrt{n}(\bar{x}-\mu)}{s}$$

$$\bar{X} = \frac{\sum X}{n} = \frac{66}{10} = \quad 6.6$$

$$S^2 = \frac{\sum x^2 - \frac{(\sum X)^2}{n}}{n-1}$$

$$1 = \frac{464 - \frac{(66)^2}{10}}{9} = 3.156$$

8	64
6	36
5	25
8	64
$\sum X$ =66	$\sum x^2$ =464

Therefore, $S = \sqrt{3.156}$= 1.78. Then, $t_{cal} = \frac{\sqrt{10}(6.6-6.0)}{1.78} = 1.07$

Hence, the corresponding table value is obtained using $t_{n-1}\left(\frac{0.05}{2}\right) = t_9(0.025) = 2.262$.

Notice the degrees of freedom 9 is the denominator of the variance formula.

Since t_{cal}= 1.07 does not fall into the critical region ($t_{cal} < t_{tab}$), H_0 should be accepted which tends to support the administrator's claim.

3.1.3 Hypothesis Testing Using P-values

The hypothesis testing techniques described about a population mean above is the classical or traditional approach. An alternative approach for hypothesis testing is the use of the probability-value, or simply the p-value. The p-value is easy, used by computer and is fast replacing the traditional approach.

The p-value is the smallest level of significance for which the observed sample information becomes significant, provided the null hypothesis is true. The p-value is computed by finding the probability that the test statistic could be the value it is or a more extreme value (in the direction of the alternative hypothesis) when the null hypothesis is true.

Note that the p-value for a hypothesis test is always a positive quantity (since it represents a probability) regardless of whether x is above or below the claimed population mean μ (or whether Z or t is positive or negative)

In general, given the significance level α, we calculate the p-value associated with the sample mean and compare these two values:

If p-value <α, then we reject the null hypothesis. If p-value >α, then we fail to reject the null hypothesis.

A Five – steps to using the p-values approach:

Step 1: State the null hypothesis (H_0) and the alternative hypothesis (H_1)

Step 2: Determine the level of significance, α

Step 3: Compute the value of observed test statistic.

Step 4: Calculate the p-value

Step 5: Determine the results as follows:

a. Compare the calculated p-value to the level of significance, □, from step 2

b. Make a decision about H_0

c. Conclude about H_1

Example 6:

Use p - values to test the following at α = 0.05. H_0: μ = 45, H_1: μ > 45, n = 100, $\bar{x}$ = 50, s = 15

Solution:

Step 1: H_0: μ = 45, H_1: μ > 45

Step 2: α= 0.05

Step 3: The calculated value of the test statistic is given as:

$$Z_{\text{cal}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{50 - 45}{15/\sqrt{100}} = \frac{5}{15/10} = 5/1.5 = 3.33.$$

Step 4: We now determine what part of the distribution relative to Z* represents the p-value.

The alternative hypothesis indicates we are interested in that part of the probability distribution that lies to the right of z*, since the "greater than" sign was used. Using the z probability table (taught to you earlier under probability), we have

p-value = $p(Z > Z_{cal}) = 1 - p(Z < Z_{cal}) = 1 - p(Z < 3.33) = 1 - 0.9996 = 0.0004$

Step 5: The p-value for this hypothesis test is 0.0004. Comparing, sincethe p-value is less than 0.05, then we reject the null hypothesis. Therefore, the mean is greater than 45 at 5% level of significance.

3.2 ONE SAMPLE TEST FOR POPULATION PROPORTION

Sometimes it might be of interest to consider the population proportion (P) instead of the population mean. The procedure for testing the hypothesis is the same except that the standard error of sample proportion is used to replace that of the sample mean.Hence, the possible hypotheses are:

$H_0: P = P_0$ against

(i) $H_1: P > P_0$

or

(ii) $H_1: P < P_0$

or

(iii) $H_1: P \neq P_0$

where P stands for the population proportion and P_0 the given or known population proportion

Therefore, the test statistic is given as $Z_{cal} = \dfrac{p - P_0}{\sqrt{\dfrac{pq}{n}}}$.

where p is the sample proportion, derived from the sample data and 'n' is the sample size. The denominator $\sqrt{\frac{pq}{n}}$ is the standard error of the sample proportion.

Example 7:

A hospital manager claims that 30 percent of his staff enjoy his leadership style. An opinion poll was conducted amongst his staff to clarify this claim. As a result 64 of his staff were chosen and 28 claim that they enjoy the lecture. Test the validity of this claim at 5 percent level of significance.

Solution:

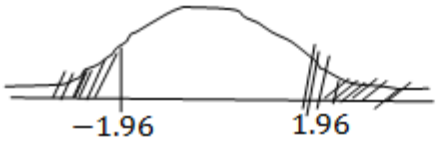
Given population proportion be $P_0 = \frac{30}{100} = 0.3$

Sample proportion = $p = \frac{28}{64} = 0.44$

That is, $H_0: P = 0.3$ against $H_1: P \neq 0.3$(two-tailed test).

The test statistics becomes $z_{cal} = \frac{p - P_0}{\sqrt{\frac{pq}{n}}} = \frac{0.44 - 0.3}{\sqrt{\frac{(0.3)(0.7)}{64}}} = 2.44$

The corresponding table value at 5% level is $Z_{tab} = Z_{0.025} = \pm 1.96$



Hence, we reject H_0 (since 2.44 > 1.96)and conclude that the teacher's claim is not valid.

3.3 TWO-SAMPLE TEST FOR MEAN AND PROPORTION

Statisticians are usually faced with the problem of comparing two samples. The aim is often to check for significant difference and relationship. Sometimes, the aim is to compare proportions. The usual assumptions is that the two samples come from normally

distributed populations and if this assumption is not sure to hold, the best and simple thing to do is to keep the sample size large. The steps for hypothesis testing remain the same as for one sample test and the determination of test statistics follows like in one sample test procedure.

3.3.1 Paired Sample T Test:

Paired sample refers to two sets of measurements of:

i. same individuals taken at two different points in timeor on two attributes, e.g.,

UME and Post-UME scores of prospective students, Weight and height measurements of pregnant women at a clinic.

ii. an attribute taken for two sets of individuals who are connected. E.g., the heights a father and that of his first son.

They are sometimes referred to as repeated measurements and always occur in pairs. The test procedure is similar to that of a one sample t test, except that the sample mean and standard deviation in the test statistic is replaced by the mean and standard deviation of the paired difference $\bar{d}$ and S_d.

Given a set of paired samples (x_i, y_i), i = 1, …, n. Let $d_i$be the difference between the ith paired observation ($d_i = x_i - y_i$). Then, $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ and $S_d = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}$.

The test statistic then becomes: $t_{cal} = \frac{\bar{d} - \mu_d}{S_d / \sqrt{n}}$, which follows a t distribution with n-1 degrees of freedom (df).

The usual null hypothesis is that there is no mean difference, H_0: $\mu_X - \mu_Y = 0$ or H_0: $\mu_X = \mu_Y$ or H_0: $\mu_d = 0$, where $\mu_d = \mu_X - \mu_Y$. The alternative could be in any of the three forms: H_1: $\mu_X - \mu_Y > 0$ or H_1: $\mu_X - \mu_Y < 0$ or H_1: $\mu_X - \mu_Y \neq 0$

Example 1:

Test the given hypothesis at $\alpha = 0.05$ level of significance using the given information: $H_0: \mu_d = 0$, $H_0: \mu_d > 0$, $n = 16$, $\bar{d} = 12$, $S_d = 22$.

Solution:

$H_0: \mu_d = 0$, $H_0: \mu_d > 0$

The calculated value is: $t_{cal} = \frac{\bar{d} - \mu_d}{S_d / \sqrt{n}} = \frac{12}{22/\sqrt{16}} = \frac{12}{22/4} = 2.18,$

The critical value t_{crit} is obtained from the table as: $t_{crit} = t_{\alpha,n-1} = t_{0.05,15} = 1.753.$

Since the calculated value is greater than the critical value, we reject H_0.

Example 2:

Given the following paired sample data, test whether there is a significant difference between X and Y at $\alpha = 0.05$.

I	X	Y
1	5	5
2	6	3
3	7	6
4	4	7
5	6	1

Solution:

I	X	Y	d = X - Y	$(d_i - \bar{d})^2$
1	5	5	0	1.44
2	6	3	3	3.24
3	7	6	1	0.04

4	4	7	-3	17.64
5	6	1	5	14.44
Total			**6**	**36.8**

H_0: $\mu_X - \mu_Y = 0$; H_1: $\mu_X - \mu_Y \neq 0$

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{6}{5} = 1.2; S_d = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1} = \frac{36.8}{5-1} = \frac{36.8}{4} = 9.2.$$

$$\therefore\ t_{cal} = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}} = \frac{1.2}{9.2/\sqrt{5}} = \frac{1.2}{9.2/2.2361} = \frac{1.2}{9.2/2.2361} = \frac{1.2}{4.1143} = 0.292,$$

To obtain the critical value, we divide α by 2, since it is a two-tailed test.

3.3.2 Test for Means of two Independent Populations with Known Variances or Large Sample

For two large population means, we consider the difference between their means, $\mu_1 - \mu_2$., the inferences made about $\mu_1 - \mu_2$ will be based on the difference between the observed sample means, $\bar{X}_1 - \bar{X}_2$.

If independent samples of sizes n_1 and n_2 are drawn randomly from large population with means μ_1, μ_2 and variances σ_1^2, σ_2^2, respectively, the sampling distribution of the difference of the sample means, $\bar{X}_1 - \bar{X}_2$;

i. is approximately normally distributed,

ii. has a mean $\mu_1 - \mu_2$,and

3. a standard error $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Since the distribution is approximately normal, we use the z – statistic. For the hypothesis tests, test statistic is:

$$Z_{\text{cal}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}};$$

if both σ_1^2 and σ_2^2 are known.

If, however, both σ_1^2 and σ_2^2 are unknown and both $n_1 > 30$ and $n_2 > 30$, we will use the test statistic:

$$Z_{\text{cal}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

Where S_1^2 and S_2^2 are the estimates of σ_1^2 and σ_2^2, respectively.

The null hypothesis states the equality of two population parameters with three possible alternatives as shown below:

$$H_0: \mu_1 = \mu_2$$

against

(i) $H_1: \mu_1 > \mu_2$

or

(ii) $H_1: \mu_1 < \mu_2$

or

(iii) $H_1: \mu_1 \neq \mu_2$

where μ_1 *and* μ_2 are the population means of populations 1 and 2 respectively.

Example 3:

A salesman states that average sales recorded from two products A and B are the same with variances 16 and 25 respectively. A consumer of one of the products arguing that the average sales from product A is greater than that of B took random samples of sizes 18

and 20 from the sales of products A and B and found their means as 40 and 36. Verify the salesman's claim at 5% level of significance.

Solution:

The null is $H_0: \mu_A = \mu_B$ against the alternative $H_1: \mu_A > \mu_B$

Since, the variances are provided, we apply,

$$Z_{cal} = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{40 - 36}{\sqrt{\frac{16}{18} + \frac{25}{20}}} = \frac{4}{\sqrt{2.1389}} = 2.735$$

for $\alpha = 0.05$, the corresponding table value is $Z_{0.05} = 1.645$

As can be seen, the calculated value is greater than the critical value. This calls for rejection of the null hypothesis.

Example 4:

Using the following information, test H_0 against H_1 at 0.05 level of significance.

$H_0: \mu_1 - \mu_2 = 0$, $H_1: \mu_1 - \mu_2 = 0$; $n_1 = 40$, $\bar{X}_1 = 24$, $S_1^2 = 7$; $n_2 = 35$, $\bar{X}_2 = 29$, $S_2^2 = 8$

Solution:

$H_0: \mu_1 - \mu_2 = 0$, $H_1: \mu_1 - \mu_2 = 0$; $\alpha = 0.05$

Although σ_1^2 and σ_2^2 are unknown, their estimates are given and the sample sizes are large. Therefore, we use a z test.

$$Z_{\text{cal}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} = \frac{(24 - 29)}{\sqrt{7/40 + 8/35}} = \frac{-5}{\sqrt{0.175 + 0.2286}} = \frac{-5}{\sqrt{0.4036}} = -7.87.$$

The critical value for this test is 1.645. Since the critical region is on the left, we see that the calculated value is less than minus the critical value, i.e., -7.87 < -1.645. So, H_0 is rejected at $\alpha = 0.05$.

3.3.3 Test for Means of two Independent Populations with Unknown Variances and Small Sample

When both population variances are not known, they can be estimated by the sample variance s^2. Then, for sample sizes $n_1 < 30$ and $n_2 < 30$, and assuming that $\sigma_1^2 \neq \sigma_2^2$,the two sample statistic is:

$$t_{\text{cal}} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \mu_0}{\sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}} \sim t_k$$

This statistic has approximately t_k distribution with $k = \min(n_1 - 1, n_2 - 1)$ **degrees of freedom.**

3.3.4 Pooled Two- Sample T Test

In the above tests, it was assumed that $\sigma_1 \neq \sigma_2$. What if there is reason to believe that $\sigma_1 = \sigma_2 = \sigma$ (even though the two are unknown). This gives us the power to pool the two samples together and estimate the standard error as:

$$S_P = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}}$$

Where s_1 and s_2 are sample variances. The t statistic is then

$$T_{\text{cal}} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Example 5:

Weight gains (in kg) of babies from birth to age one year are measured. All babies weighed approximately the same at birth.

Group A:5 7 8 9 6 7 10 8 6
Group B:9 10 8 6 8 7 9

Assume that the samples are randomly selected from independent normal populations. Is there any difference between the true means of the two groups?

i) Assume $\sigma_1 = \sigma_2 = 1.5$ is known.

ii) Assume σ_1 *and* σ_2 are unknown and unequal.

iii) Assume σ_1 *and* σ_2 are unknown but equal.

Solution:

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$

Where μ_1 is the true population mean of the group A and μ_2 is the true population mean of group B.

$\bar{x}_1 = 7.33$

$s_1 = 1.58$ $\bar{x}_2 = 7.33$

$n_1 = 9$ $s_2 = 1.58$

$n_2 = 7$

i) $\sigma_1 = \sigma_2 = 1.5$**.** Then, the two sample statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{7.33 - 8.14}{1.5 \times \sqrt{\frac{1}{9} + \frac{1}{7}}} = -1.07$$

The two sided P- value is

$$2P(Z \geq |z|) = 2P(Z \geq 1.07) = 0.28$$

Where Z ~ N(0, 1).

So there is no difference between the true population mean of these two groups at the significance level 0.1.

A 90% confidence interval for $\mu_1 - \mu_2$ is:

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm Z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(7.33 - 8.14) \pm 1.645 \times 1.5 \sqrt{\frac{1}{9} + \frac{1}{7}}$$

$$= (-2.05, 0.43)$$

This is consistent with rejecting H_0 at significance level 0.1 in the two sided Z - test.

ii) Assuming σ_1 and σ_2 are unknown and unequal. Then, the test statistic is

$$t = \frac{\bar{x} - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s^2{}_2}{n_2}}} = \frac{7.33 - 8.14}{\sqrt{\frac{1.58^2}{9} + \frac{1.35^2}{7}}} = -1.10$$

The two-sided p-value is

$$2P(T \geq |t)|) = 2P(T \geq 1.10) = 0.30$$

$T \sim t_6$

where

A 90% CI for $\mu_1 - \mu_2$ is given by

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= (7.33 - 8.14) \pm 1.94 \times \sqrt{\frac{1.58^2}{9} + \frac{1.35^2}{7}}$$

$$= (-2.23, 0.61)$$

Where $P(|T| < t^*) = .0.90$. That is

$P(T > t^*) = 0.05$

iii) Assume σ_1 and σ_2 are unknown but equal.

The pooled two- sample estimator of σ is

$$S_P = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}} = \sqrt{\frac{(9-1) \times 1.58^2 + (7-1) \times 1.35^2}{9 + 7 - 2}} = 1.54$$

Thus, the pooled two –sample *t* statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{7.33 - 8.14}{1.54\sqrt{\frac{1}{9} + \frac{1}{7}}} = -1.04$$

The two-sided p-value is given by

$2P(T \geq |t)|) = 2P(T \geq 1.04) = 0.32$

Where $T \sim t_{14}$

A 90% CI for $\mu_1 - \mu_2$ is given by

$$\cdot (\bar{x}_1 - \bar{x}_2) \pm t^* \times S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= 7.33 - 8.14 \pm 1.77 \times 1.54\sqrt{\frac{1}{9} + \frac{1}{7}}$$

= (-2.18, 0.56)

Where $P(|T| < t^*) = .0.90$. That is, $P(T > t^*) = 0.05$

3.3.5 Test for the Equality of Two Population Proportions

The null hypothesis in this case states the equality of two population proportions with three possible alternative hypotheses as shown below:

$H_0: P_1 = P_2$

against the alternatives

(i) $H_1: P_1 > P_2$

or

(ii) $H_1: P_1 < P_2$

or

(iii) $H_1: P_1 \neq P_2$

where P_1 and P_2 are proportions from populations 1 and 2 respectively.

The test statistics is thereby given as

$$Z_{cal} = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

$p_1$and $p_2$are the proportions derived from sample drawn from population 1 and 2 respectively.

Example 6:

The manager of a drug manufacturing company which produces two drugs (A and B) claims that the proportion of consumers purchasing drug A is the same as the proportion purchasing drug B. The Head of marketing unit of the company doubting this claim collected samples of sales records of 36 and 25 potential consumers from the locality and found out that 24 and 15 purchase drugs A and B respectively. Can you support the manager's claim at 5 percent level of significance?

Solution:

$H_0 : P_A = P_B$ against $H_1 : P_A \neq P_B$

$n_A = 36$; $n_B = 25$, $\bar{p}_A = \frac{24}{36} = 0.667$; $\bar{p}_A = \frac{15}{25} = 0.600$

Now, the test statistic is $Z_{cal} = \frac{p_A - p_B}{\sqrt{\frac{pAqA}{n_A} + \frac{pAqA}{n_B}}} = \frac{0.667 - 0.6}{\sqrt{\frac{(0.667)(0.333)}{36} + \frac{(0.6)(0.4)}{25}}} = \frac{0.067}{0.126} = 0.532$

For $\alpha = 0.05$, the corresponding table value is $Z_{tab} = Z_{0.025} = \pm 1.96$. H_0 is accepted and conclude that the manager's claim is valid.

4.0 CONCLUSION

In this unit, we discussed in details, various procedures for testing hypothesis. Several examples were used to demonstrate the procedures to allow easy understanding for health practitioners and researchers.

5.0 SUMMARY

Several statistical methods for different hypothesis testing have been presented. This includes t test and z test for testing hypothesis involving single population mean and proportion. The case of two sample test was also considered.

6.0 TUTOR-MARKED ASSIGNMENT

1. A research executive wishes to test the hypothesis that the mean family income in a particular area is more that ₦50,000 per month. The standard deviation of the population of monthly income in the area is known to be ₦3,500. He then collected data

from 144 families, randomly selected from the area and found the average monthly income as ₦50,500.

(a) State the null and alternative hypotheses.

(b) Carry out a test of the null hypothesis using 5 percent level of significance.

2. A researcher observed that three out of every fifteen of their products are normally defective. A total of 200 samples of the products were being tested. If the sample is normally distributed and 45 of the products were identified to be defective, test the hypothesis that the observation of the researcher is true at 5 percent level of significance.

3. A lecturer claims that the proportion of students passing course A is the same as the proportion of students passing course B. The course coordinator doubting this claim collected samples of scores of 16 and 25 students from the population who sat for course A and course B and found out that 12 and 15 passed the courses respectively. Can you support the lecturer's claim at 5 percent level of significance?

4. A trader claims that on the average, his daily sale is ₦58. A researcher trying to verify this claim, took a random sample of sales for 10 days and recorded them as

60, 56, 35, 74, 52, 63, 59, 55, 70.

Can you conclude that the traders claim is valid at 5% level of significance?

7.0 REFERENCES/FURTHER READING

i. Mojekwu J. N. (2012). Business Statistics with Solved Examples, Easy Print Publication, Lagos.

ii. Onyeka-Ubaka, J. N. (2013). *Multi-Level Statistics: An Academic Companion for Inter-Disciplinary Professional Competence*, Royal Choice Multi-Media, Lagos.

ii. Spiegel, A. (2011). Statistics. New York, Schum Series.

iii. Rose, B. (1995). Fundamentals of Biostatistics (4th ed.), Duxbury Press, ITP, USA.

iv. Le, C. T. (2003). Introductory Biostatistics. John Wiley and Sons Publishing, Wiley-Interscience. Department of mathematics, University of Lagos (2015). A First Course in Statistics. Lagos: Nile Ventures.

UNIT 3 ONE-WAY ANOVA AND CHI SQUARE TEST

CONTENT

1.0 INTRODUCTION

In this unit, we shall treat in detail one way analysis of variance as a statistical procedure to compare means of more than two samples. The chi square test for analyzing categorical data will also be presented. Rigorous derivations shall be omitted but illustrations with examples will be presented.

2.0 OBJECTIVES

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

i. Explain the procedure for one way ANOVA
ii. Formulate hypothesis for one way ANOVA
iii. Test hypothesis to compare means of multiple samples
iv. Carry out chi square test of hypothesis for goodness of fit
v. Test hypothesis for association between two categorical variables

3.0 MAIN CONTENT

3.1 ANALYSIS OF VARIANCE (ANOVA)

Analysis of variance is a method for splitting the total variation of our data into meaningful components that measure different sources of variation. It is used to test for the equality of several means simultaneously. This comparison actually involves analyzing variances, which are scaled sums of squares reflecting different sources of variability. ANOVA is similar to independent *t* test but for more than two samples. The procedure for the analysis of variance to be considered in this section will omit the derivations.

The test is an F-tests, which is a ratio of two variances. These sums of squares are constructed so that the statistic tends to be greater when the null hypothesis is not true. In

order for the statistic to follow the F-distribution under the null hypothesis, the data should be statistically independent, normally distributed and have equal variance.

3.1.1 The One-way ANOVA Formulation

The simplest form of ANOVA is the one way ANOVA, so called because only one source of variation is of interest. That is there is only one independent variable, which must be categorical or in groups against a continuous measurable response or dependent variable. The independent variable is always referred to as the treatment and the categories as treatment levels.

The model of a one way ANOVA is given by:

$$y_{ij} = \mu + \alpha_j + e_{ij}; \quad i = 1,2,\cdots,n_j;\ j = 1,2,\cdots,t,$$

where,

y_{ij} = the *i*th observation on the *j*th treatment.

μ = the grand mean

α_j = the mean effect of the*j*th treatment category

e_{ij} = error term.

Consider the data layout for a one-way ANOVA below, where the data were collected on a response variable *Y*.

	Treatment				
	1	2 ⋯	j ⋯	t	Grand
	Y_{11}	Y_{12} ⋯	Y_{1j} ⋯	Y_{1t}	
	Y_{21}	Y_{22} ⋯	Y_{2j} ⋯	Y_{2t}	
	⋮	⋮	⋮	⋮	
	Y_{i1}	Y_{i2} ⋯	Y_{ij} ⋯	Y_{it}	
	⋮	⋮	⋮	⋮	

	$Y_{n_1 1}$	$Y_{n_2 2}$	$Y_{n_j j}$	$Y_{n_t t}$	
Totals	T_1	T_2	T_j	T_t	T
Number	n_1	n_2	n_j	n_t	N
Means	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_j$	$\bar{Y}_t$	$\bar{Y}$

where T_j is the totals of observations under *j*th treatment level, n_j is the number of observations in *j*th treatment group, $\bar{Y}_j$ is the mean of observations in *j*th treatment level, T is the grand total of all observations, $\bar{Y}$ is the grand mean of all observations and *N* is the total number of observations.

Sums of squares

Omitting all derivations, we state the following.

Variability is expressed as a sum of squares. The total variability in the dependent variable Y is partitioned into variability due to the treatment variability due to random error. That is,

Total sum of squares = Treatment sum of squares + Error sum of squares.

$\therefore TSS = TrtSS + ESS$.

For computational convenience, the following formulae are usually preferred.

$TSS = \sum_{j=1}^{t}\sum_{i=1}^{n_j} Y_{ij} - \frac{Y_{..}^2}{N}$, with *N*-1 degrees of freedom;

$TrtSS = \sum_{j=1}^{t} \frac{Y_{.j}^2}{n_j} - \frac{Y_{..}^2}{N}$, with *t*-1 degrees of freedom;

$ESS = \sum_{j=1}^{t}\sum_{i=1}^{n_j} Y_{ij}^2 - \frac{Y_{.j}^2}{n}$, with *N*-*t* degrees of freedom.

It is much easier to obtain ESS by subtraction as: ESS = TSS – TrtSS

Hypothesis and Test Statistics for the *F* test of One Way ANOVA

The usual null hypothesis in that there is no significant difference in the treatment means or no significant treatment effect while alternative hypothesis states that at least one treatment mean is significantly different from others or the treatments have significant effect on the response variables. That is:

Test the hypothesis,

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \cdots = \mu_k$

H_1: $\mu_i \neq \mu_j$, for at least one $i \neq j$.

Or,

H_0: There is no significant treatment effect; ie, $\tau_i = 0$.

H_1:There is a significant treatment effect; ie, $\tau_i \neq 0$.

The mean squares are defined as follows:

Total Means Square, TMS = TSS/(*N*-1)

Treatment Mean Square, TrtMS = TrtSS/(*t*-1)

Error Mean Squares, EMS = ESS/(*N*-*t*).

Under certain conditions, the F test statistic is the ratio: $F_{cal} = \frac{TrtMS}{EMS} \sim F_{\alpha,(v1,v2)}$, where $v_1 = t - 1$ is the treatment degrees of freedom and $v_2 = N - t$ is the error degrees of freedom.

Decision Rule: The decision rule is to reject the null hypothesis if $F_{cal} > F_{\alpha,(v1,v2)}$ at a given significance level α.

ANOVA Table:

Source of variation	Sum of Squares (SS)	Degrees of freedom (df)	Mean Squares (MS)	F_{cal}	$F_{crit.}$
Treatment	TrtSS	t-1	TrtSS/t-1 = TrtMS	TrtMS/EMS	$F_{\alpha,(v1,v2)}$
Error	ESS	N-t	ESS/N-t = EMS		
Total	TSS	N-1	TSS/N-1		

Example 1:

Mean liver weight (expressed as a percentage of body weight) of rats from 4 Groups A, B, B, C, D are presented below.

A	B	C	D
3.40	3.37	3.34	4.64
3.96	3.67	3.75	3.93
5.89	3.34	3.81	3.77
4.19	3.47	3.66	4.81
3.63	3.39	3.55	4.21
3.76	3.41	3.51	3.88
3.84	3.55		3.96
	3.44		3.91

Let μ_i be the mean liver weight for group *i*.

Test the hypothesis H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ against Hi: $\mu_i \neq \mu_j$ for some i, j at 5% level of significance.

Solution:

No.	A	B	C	D
1	3.40	3.37	3.34	4.64
2	3.96	3.67	3.75	3.93
3	5.89	3.34	3.81	3.77
4	4.19	3.47	3.66	4.81
5	3.63	3.39	3.55	4.21
6	3.76	3.41	3.51	3.88
7	3.84	3.55		3.96
8		3.44		3.94
$T_{.j}$	28.67	27.64	21.62	33.11	$T_{..} = 111.04$
$\bar{Y}_{.j}$	4.096	3.455	3.603	4.139	$C.F. = \frac{T_{..}^2}{N}$
$\sum Y_{ij}^2$	121.5499	95.5786	78.0524	138.0717	$\sum_i \sum_j Y_{ij}^2 = 433.2526$

Where $n_A = 7 \quad n_B = 8 \quad n_C = 6 \quad n_D = 8 \quad n = n_A + n_B + n_C + n_D$

$= 29$

$$C.F. = \frac{T_{..}^2}{N} = \frac{(111.04)^2}{29} = 425.17.$$

The total sum of squares, $SS_{Total} = 3.40^2 + 3.96^2 + 5.89^2 + ... + 3.96^2 + 3.91^2 - \frac{(111.04)^2}{29}$

$= 8.0843.$

Treatment Sum of squares,

$$SS_{Treatment} = \frac{(28.67)^2}{7} + \frac{(27.64)^2}{8} + \frac{(21.62)^2}{6} + \frac{(33.11)^2}{8} - \frac{(111.04)^2}{29} = 2.6901$$

ANOVA TABLE

Source	Sum of squares	df	MS	F
Between group	2.6901	3	0.8967	4.16
Within Group	5.3942	25	0.2158	
TOTAL	8.0843	28		

Conclusion: Since $F_{calculated} = 4.16$ and $F_{table} = 2.99$ at $\alpha = 0.05$, $F_{calculated}$ is greater than F_{table} , we reject null hypothesis and conclude that the mean liver weight for the groups are not equal for at least one group.

Example 2:

A company supplies a customer with a larger number of batches of raw materials. The customer makes three sample denominations from each of five randomly selected batches to control the quality of the incoming material. The data is given below:

Batch				
1	2	3	4	5
74	68	75	72	79
76	71	77	74	81
75	72	77	73	79

i. Is there a significant mean difference among the batches?

ii. Estimate the variance components.

Solution:

H_0: There is no significant mean difference among the batches.

H_1: There is a significant mean difference among the batches.

Following the ANOVA procedure, we do the following:

$$SS_{BG} = \frac{225^2 + 211^2 + 229^2 + 219^2 + 239^2}{3} - \frac{1123^2}{15},$$

$$= \frac{50625 + 44521 + 52441 + 47961 + 57121}{3} - \frac{1261129}{15},$$

$$\therefore SS_{BG} = \frac{252669}{3} - \frac{1261129}{15} = 84223 - 84075.27 = 147.73,$$

$$SS_{Total} = 74^2 + 68^2 + 75^2 + \cdots + 73^2 + 79^2 - \frac{1123}{15},$$

$$= 84241 - 84075.27 = 165.73,$$

$$SS_{WG} = SS_{Total} - SS_{BG} = 165.73 - 147.73 = 18,$$

ANOVA Table:

Source	***SS***	***df***	***MS***	F_{cal}	F_{crit}
Between Batches (Treatment)	147.73	4	36.93	20.52	3.478
Within Treatment (Error)	18.00	10	1.8		
Total	165.73	14			

Since F_{cal} is greater than the critical F, H_0 is rejected at 0.05 level of significance. We conclude that there is no significant mean difference among the batches.

3.2 CHI SQUARE TESTS

Categorical data is data in which each respondent belongs to one and only one category. Examples are data on Sex (male or female), Marital status (married, single, divorced, separated, widowed), Religion – (Islam, Christianity, Traditional, Jewish or others), Ethnicity (Hausa, Yoruba, Ibo, others), etc. Non-categorical data can be categorized by grouping or scaling. Example, data on age is non-categorical but when scaled to age groups, it becomes categorical data.

Categorical data skewed and cannot follow the normality assumption for most tests discussed above. Therefore, analysis of such data is done using the chi-square (χ^2) distribution. Chi-square distribution is a skewed distribution that depends on its degrees of freedom, which is its only parameter. It is important to note that sums of squares are known to follow the chi-square distribution and hence its usefulness in chi square test, since chi-square test statistics are scaled sums of squares.

Analysis of categorical and skewed data is very important because it provides basis for decision making regarding population of such data. A number of tests of hypothesis can be done for categorical data. In this study material, only the chi square test of goodness of fit and test of association will be treated.

3.2.1 Chi Square Goodness of Fit

Goodness of fit as the name implies, is used to test whether a given set of data come from a specified distribution. The test is a test of agreement between the observed frequencies and the expected frequencies. It makes use of χ^2-distribution; the degree of freedom is usually (k-1) as there is only one row or one column with k distinct groups.

The test statistics is given by:

$$\chi^2_{\text{cal}} = \sum \frac{(O-E)^2}{E},$$

where; O = observed frequencies

E = expected frequencies

The test value is to be compared with the critical values of the chi square distribution at (k-1) degrees of freedom for a given significance level (χ^2_{crit}); k is the number of distinct groups. If χ^2_{cal} is greater than $\chi^2_{\text{crit.}}$, reject H_0; where H_0 states that the given sample is from the specified distribution.

Example 1:

A manager of a big hospital which has five branches believed that the amount of revenue (₦'000) generated from each branch are the same. A survey was then carried out to see whether this claim is valid at 5 percent level of significance and the following results were obtained.

Branch	A	B	C	D	E
Revenue	180	250	230	190	150

Solution:

The null hypothesis is,

H_0: The distribution of revenue is normal (the same)

H_1: The distribution of revenue is not normal (has changed)

The assumption is that if revenue generation has not changed, then the branches are expected to record equal revenue. That is, the total revenue generated divided by the number of branches: $\frac{₦1000}{5} = ₦200$ is expected from each branch.

Branch	Observed Revenue (O)	Expected Revenue (E)	$\sum \frac{(O-E)^2}{E}$
A	180	200	2.0
B	250	200	12.5
C	230	200	4.5
D	190	200	0.5
E	150	200	12.5
	1000		32.0

$$\chi^2_{cal} = \sum \frac{(O-E)^2}{E} = 32.0 \quad \text{and} \quad \chi^2_{tab} = \chi^2_{k-1}(0.05) = \chi^2_{4(0.05)} = 9.49$$

Since the calculated statistic is greater than the critical value, there is enough evidence to reject H_0. Therefore, we conclude that the revenue generation pattern has changed or the revenue generated from the branches are not normally distributed.

Example 2:

Blood types in the general population are distributed as follows:20% have type O, 45% have type A, 20% have type B and 15% have type AB. A group of 300 people from a certain area are surveyed and found to have the following frequencies for the different blood types.

Blood Type	O	A	B	AB
Frequency	75	50	90	85

Based on the results above, can we conclude that people from this area have a different distribution of blood types at 5% level of significance?

Solution:

H_0: There is no difference in the distribution of blood type

H_1: There is a difference in the distribution of blood type

Blood Type	Observed frequency (O)	Expected frequency (E)	O - E	$(O - E)^2$	$\sum \frac{(O-E)^2}{E}$
O	75	20% of 300 = 60	15	225	3.75
A	50	45% of 300 = 135	-85	7225	53.52
B	90	20% of 300 = 60	30	900	15
AB	85	15% of 300 = 45	40	1600	35.56
Total	**300**	**300**			**107.83**

$\alpha = 0.05$, k = 4, df = k-1 = 4-1 = 3

$\chi^2_{cal} = 107.83;\ \chi^2_{0.05,(3)} = 7.82.$

Since the calculated value is greater than the critical value, reject H_0 and conclude that there is a difference in the distribution of blood type of people in the area from the general population.

3.2.2 Chi Square Test of Association

Sometimes, statisticians are faced with the problem of testing whether there is any significant association between two categorical variables. For example, a researcher might wish to find out whether a patient's nutritional class is related to their blood pressure. The assumption here is that if the a patient's nutritional class is independent of their blood pressure then the proportion of patients with a particular nutritional should be equal to the proportion of patients with a particular blood pressure.

The test of association is a chi square test involving a contingency table. A contingency table is an array of data in a two-way dimension of rows and columns. Consider the following contingency table:

		Variable 1 (X)				Row Totals
		C_1	C_2	. . . . .C_j . . .	C_c	
Variable 2 (Y)	R_1	O_{11}	O_{12}	. . . . .	O_{1c}	RT_1
	R_2	O_{21}	O_{22}	. . . . .	O_{2c}	RT_2
	.	.	.			.
	R_i	.	.O_{ij}			RT_i
	.	.	.			.
	R_r	O_{r1}	O_{r2}	. . . . .	O_{rc}	RTr
Column Totals		CT_1	CT_2		CT_jCT_c	GT

Where

X is a categorical variable (variable 1) with columns C_1 to C_c

Y is another categorical variable (variable 2) with rows R_1 to R_r

c = the number of columns of X

r = the number of rows of Y

O_{ij} = observed frequency forrow i and column j.

RT_i = Total for row i

CT_j = Total for column j.

GT = Grand Total

The null hypothesis is that the two categorical variables are not related or they are independent against appropriate alternative.

The test statistic is:

$$\chi^2_{cal} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}};$$

where

Eij is the expected frequency computed as:

$$E_{ij} = \frac{RT_i \times CT_j}{GT}.$$

The expected values are calculated based on the assumption that if there is no association between the two variables, then the proportion each cell is contributing to the grand total would be equal.

The degrees of freedom for the test is (r-1)(c-1).

The test value is then compared with the critical values of the chi square distribution at (r-1)(c-1) degrees of freedom for a given significance level (χ^2_{crit}). If χ^2_{cal} is greater than $\chi^2_{crit.}$, reject H_0.

Example 3:

A research executive intends to verify the general belief expressed by the public that there is no association between one's age and the type of food he likes to eat. A sample of

250 people were administered with questionnaires and the responses were obtained as given below

Age of Individuals Classified by Food

Age \ Food	Rice	Eba	Yam
Adult	35	45	40
Children	30	55	45

(i) State the appropriate hypotheses for the test

(ii) Carry out the test at 5% level of significance

(iii) State clearly your decision and conclusion.

Solution:

(i) H_0: Association does not exist between age of an individual and the type of food he likes to eat.

H_1: Dependency exists between the two cross-classified variables.

(ii) The test statistic is given as

$$\chi^2_{cal} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} .$$

To obtain the expected frequencies we use $E_{ij} = \frac{RT_i \times CT_j}{GT}$. For example, to obtain the expected frequency for observed value in row 1, column 1, we have $E_{11} = \frac{120 \times 65}{250} = 31.2.$

Now, we can rearrange the data on table 10.3 as follows:

Observed (O_{ij})	Expected (E_{ij})	$\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

35	31.2	0.4628
30	33.8	0.4272
45	48	0.1875
55	52	0.1731
40	40.8	0.0157
45	44.2	0.0145
		1.287

Hence, to get the corresponding hypothesized table value, we read χ^2 – distribution table at

(2-1)(3-1) = 2 degree of freedom for the given level of significance. That is,

$$\chi^2_{0.05,(2)} = 5.99.$$

This calls for acceptance of H_0, since the table value is more than the calculated value.

4.0 CONCLUSION

In this unit, we discussed in detail the one way ANOVA procedure for testing the means of multiple samples. Chi square tests of goodness of fit and association have been considered too. Several examples were used to illustrate both procedures.

5.0 SUMMARY

One-way ANOVA is a statistical method for testing hypothesis on multiple samples. It is used to compare means for more than two groups by analysing variances. On the other hand, the chi square test is used to test hypothesis for categorical variables.

6.0 TUTOR-MARKED ASSIGNMENT

1. A researcher was interested in determining whether significant influence of the salary structure on performance of workers exists in an organisation. He obtained the salaries from four groups of workers whose employers use different salary structure.

Group A	Group B	Group C	Group D
4	3	10	8
5	8	13	5
7	6	4	10
9	4	5	6
10	2	8	4

(a) What type of statistical analytical technique is required to execute this task?

(b) Formulate appropriate hypotheses and test at 5 percent level of significance.

2. A researcher intends to verify the general belief expressed by the public that there is no dependency between one's genotype status and protein intake. A sample of 400 people administered with questionnaire and the responses were obtained as given below.

Group / Protein Intake	AA	AS	SS
High	88	90	65
Moderate	40	7	20
Low	10	48	32

(a) State the appropriate hypotheses.

(b) Carry out the test at 5 percent level of significance.

7.0 REFERENCES/FURTHER READING

i. Mojekwu J. N. (2012). Business Statistics with Solved Examples, Easy Print Publication, Lagos.

2. Onyeka-Ubaka, J. N. (2013). *Multi-Level Statistics: An Academic Companion for Inter-Disciplinary Professional Competence*, Royal Choice Multi-Media, Lagos.

3. Spiegel, A. (2011). Statistics. New York, Schum Series.

4. Rose, B. (1995). Fundamentals of Biostatistics (4^{th} ed.), Duxbury Press, ITP, USA.

5. Le, C. T. (2003). Introductory Biostatistics. John Wiley and Sons Publishing, Wiley-Interscience. Department of mathematics, University of Lagos (2015). A First Course in Statistics. Lagos: Nile Ventures.

UNIT 4 CORRELATION AND REGRESSION

CONTENT

1.0 INTRODUCTION

When two or more factors tend to vary simultaneously, the question is which of these factors vary with others and to what extent do they vary.

Correlation and regression are two techniques which enable us to see the relationship between the actual dimensions of two or more variables. It is important to understand what these two techniques have in common and the differences between them. While Regression is concerned with the way in representing the relationship between these variables correlation only measures the strength of the relationship.

Regression is therefore, a statistical tool that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the other. On the other hand, correlation models are used to study the nature of the relations between the variables, and also may be used for making inferences about any one of the variables on the basis of the others.

One difference between correlation and regression is that there is dependent relationship in regression, but correlation need not be dependent. Furthermore, factors that are correlated need not show significant regression but factors that show significant regression are always correlated.

2.0 OBJECTIVES

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

i. Explain correlation and regression

ii. Compute different correlation coefficients

iii. Test hypothesis concerning population coefficients

iv. Formulate a simple linear regression model and estimate its parameters

v. Test hypothesis concerning regression parameters

vi. Formulate a multiple linear regression model

vii Apply models in prediction.

MAIN CONTENTS

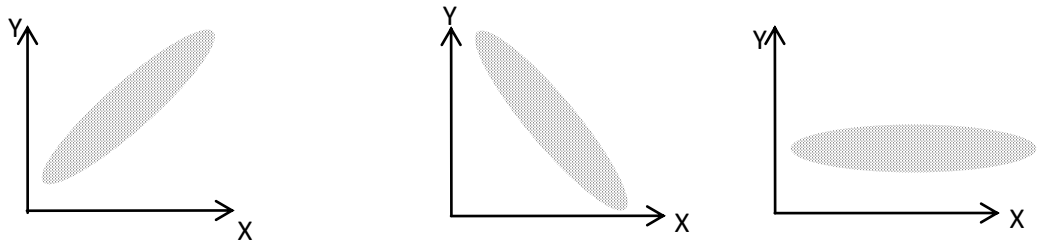
3.1 CORRELATIONANALYSIS

Correlation is a statistical tool used to study relationship between two or more variables. If two variables are involved, it said to be bivariate. If more than two variables are involved, it is said to be multivariate correlation, in which case the correlations between any two pairs is called partial correlations.

3.1.1 Bivariate Data and Scatter Diagram

Bivariate data (x,y) is a set of values which appears in pairs. Sometimes, the value of one variable (y) depends on the value of the other (x). It is of interest how they covary.

Scatter diagram is the graphical display of bivariate data by plotting y_i values on the y-axis and the x_i values on the x-axis. It is used to illustrate whether a relationship exist between any two variables and to indicate the kind of relationship. The diagrams below give illustration of different types of correlation.



Positive Correlation Negative Correlation No Correlation

Positive Correlation: Both variables vary in the same direction, that is, when one increase, the other increase or when one decrease, the other decrease.

Negative Correlation: Both variables vary in opposite directions, that is, when one increase, the other decrease.

Zero Correlation: Both variables do not covary.

3.1.2 Correlation Coefficients

This is the statistical measure, which determines the amount of linear relationship that exists between two variables. There are three types of correlation coefficient:

i. Karl Pearson's correlation coefficient

ii Spearman rank correlation coefficient

iii Kendall's correlation coefficient.

We shall only be discussing the first two in the course of this text.

Karl Pearson's Correlation Coefficient (ρ)

This is used for any bivariate population, which are normally distributed. The value of ρ always ranges from -1 to +1. It is The sample estimate r_p of the Karl Pearson coefficient is given as:

$$r_p = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(y\right)^2\right]}} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}};$$

Where, $S_{XY} = \sum (X - \bar{X})(Y - \bar{Y})$, $S_{XX} = \sqrt{\sum (X - \bar{X})^2}$ and $S_{YY} = \sqrt{\sum (Y - \bar{Y})^2}$

If r is positive, the variables are positively correlated, if r is negative, the variables are negatively correlated, while zero indicates no correlation. n is the number of paired observations.

Spearman's Rank Correlation Coefficient

It is sometimes difficult to quantify a data or if the set of data has big values, rank correlation coefficient is used to determine the extent to which the variables are related. It is also used in situation where the bivariate data involves some form of rankings.

The spearman's rank correlation coefficient r_s is defined as

$$r_s = 1 - \frac{6\sum D^2}{n\left(n^2 - 1\right)}$$

Where, D = Difference in ranks

n = no of paired observations.

The normality assumption is weakened in Spearman's correlation coefficient since it is non parametric. Therefore, only the Pearson's correlation coefficient will be used for inference in hypothesis testing.

Note the following also:

When the correlation coefficient between X and Y is 0, it implies that X and Y are independent.

Spearman's correlation coefficient is always greater than or equal to Pearson's correlation coefficient.

Partial Correlation Coefficient

When there are more than two variables in a model, partial correlation coefficient gives the correlation of two variables while controlling for a third or more other variables. Just as the simple correlation coefficient between y and x describes their joint behaviour, the partial correlation describes the behaviour of y and, say, x_1, when $x_2,\ldots,x_p$ are held fixed. The formula is omitted here.

3.1.3 Test Concerning Correlation Coefficient, ρ

Hypothesis

The usual hypothesis is: H_0: $\rho = 0$ against H_1: $\rho \neq 0$, which is essentially a two-tailed test.

The test Statistics is given by

$$t_{cal} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where r is the Pearson's correlation coefficient.

t_{cal} is to be compared the critical value of the t distribution with n-2 degrees of freedom and large values of t_{cal} leads to rejection of the null hypothesis. That is, we reject null hypothesis when $t_{cal} > t_{\alpha/2}$ ($t_{\alpha/2}$ is the critical or tabulated value of t).

Example 1:

The marks obtained by 10 students in theory and practical test papers in Physics are given below.

Student	A	B	C	D	E	F	G	H	I	J
Theory marks	79	63	84	46	77	73	56	58	49	69
Practical marks	56	42	59	35	54	62	47	51	24	49

Compute the spearman's correlation coefficient and comment on your result.

Solution:

I	Theory mks (x)	Practical mks (y)	r_x	r_y	D	d^2_i
1	79	56	9	8	1	1
2	63	46	5	3	2	4
3	84	59	10	9	1	1
4	46	35	1	2	-1	1
5	77	54	8	7	1	1
6	73	62	7	10	-3	9
7	56	47	3	4	-1	1
8	58	51	4	6	-2	4
9	49	24	2	1	1	1
10	69	49	6	5	1	1
Total	654	483				24

Where r_x and r_y are the rankings of x and y respectively in ascending order

Using

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

where, n = 10,

$$\therefore \quad r = 1 - \frac{6(24)}{10\left(10^2 - 1\right)} \quad \Rightarrow \quad r = 1 - 0.145$$

$$\therefore r = 0.85$$

Comment:

The r = 0.85 shows a strong, positive but imperfect correlation between the performance of students in the practical test and theory test, i.e., a student that did well in the practical test also did well in the theory test.

Example 2:

Use the data in example 1.6.1 to calculate the Pearson's correlation coefficient and test the hypothesis of no correlation between x and y at 5% significance level.

Solution:

Using data in example 1.6.1

I	Theory mks (x)	Practical mks (y)	x^2	y^2	Xy
1	79	56	6241	3136	4424
2	63	46	3969	2116	2898
3	84	59	7056	3481	4956
4	46	35	2116	1225	1610
5	77	54	5929	2916	4158
6	73	62	5329	3844	4526
7	56	47	3136	2209	2632
8	58	51	3364	2601	2958
9	49	24	2401	576	1176
10	69	49	4761	2401	3381
Total	654	483	44302	24505	32719

$$\sum x = 654, \sum y = 483, \ \sum x^2 = 44302, \ \sum y^2 = 24505, \ \sum xy = 32719, \ n = 10$$

$\therefore$ The Pearson' s correlation coefficient is

$$r = \frac{n\sum xy - \sum x \sum y}{\left[n\sum x^2 - \left[\sum x\right]^2\right]\left[\sum y^2 - \left(\sum y\right)^2\right.}$$

$$= \frac{10(32719) - (654)(483)}{\sqrt{\left[10(44302) - (654)^2\right]\left[10(24505) - (483)^2\right]}} = 0.84$$

The hypothesis of no correlation implies H_0: $\rho = 0$ against H_1: $\rho \neq 0$.

To test this hypothesis, we use the t statistic as given in (1.2). Thus,

$$t_{cal} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.85(\sqrt{8})}{\sqrt{1-(0.85)^2}} = \frac{0.85(2.824)}{\sqrt{1-0.7225}}$$

$$= \frac{2.4004}{0.2775} = 8.65$$

We check up the tabulated values of the t-distribution at 8 (10 - 2) degrees of freedom for $\alpha = 0.05/2 = 0.025$. Thus, $t_{0.025(8)} = 2.306$.

Since $t_{cal} = 8.65 > t_{0.025(8)} = 2.306$, test is significant and we reject H_0.

Conclusion: There is a significant correlation between x and y.

3.2 SIMPLE LINEAR REGRESSION

Regression is defined as dependent relationship between two or more variables. Therefore, regression analysis can be defined as a statistical technique used in studying a relationship between two or more variables. Most of the work in regression is to build a mathematical relationship formula that can be used to explain the dependent relationship. There is always a dependent variable, also known as response variable or outcome variable and one or more independent variables, also known as explanatory variables or predictor variables or simply predictors.

A regression model can be linear or non-linear. It is “linear” if the parameters do not appear as an exponent or multiplied or divided by another parameter and the independent variable appears only in the first power or powers that can be made linear with simple transformation. A linear regression can be simple or multiple. It is simple if there is only one independent variable, otherwise, it will be called multiple linear regression.

Uses of Regression

Regression is perhaps one of the most widely used and applied statistical techniques. It can be used by industrialists, economists, medical workers, social workers, environmentalist, government, etc. Its strong point is that data collected in the past or present can be used to build a mathematical model, which can be used to assess and control the future. Therefore, the main uses of regression can be summarized into three main headings:

Prediction

Planning

Control

3.2.1 The Simple Linear Regression Model

The General model of a simple linear regression is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where:

y_i is the *ith* value of the response variable or the dependent variable.

β_0 and β_1 are unknown parameters to be estimated.

x_i is the *ith* value of the independent variable or the explanatory variable.

ε_i is a random error term.

The parameters β_0 and β_1 in the regression model are called regression coefficients. β_1 is the slope of the regression line. It indicates the change in the mean of of y per unit change in x. The parameter β_0 is the y intercept of the regression line; if β_0 is equal to 0 then the regression passes through the origin.

To estimate the simple linear regression model above means finding estimates, say a and b respectively, for the regression parameters β_0 and β_1. This will give the equation of the regression line:

$y_i = a + bx_i$.

Assumptions of Linear Regression Models

For the purpose of both estimation and inference, the following assumptions must hold for regression models.

i. The relationship between the dependent and independent variables must be linear.
ii. The mean of the errors must be zero, ie, $E(\varepsilon) = 0$.
iii. The error variance σ^2 must be constant, a condition termed homoscedacity. (Non-constancy of the error variance is termed heteroscedacity).
iv. The error must be independent. This implies that $cov(\varepsilon_i,\varepsilon_j)$, for all i,j and $i \neq j$.
v. All the independent variables must be fixed.
vi. Errors must be normally distributed. This implies that $\varepsilon_i \sim iidN(0, \sigma_\varepsilon^2)$. Although the normality assumption is weak when estimating with least squares method, it is strongly required when making inference and estimating by maximum likelihood method.

The Distribution of Y

Consider the model: $y_i = (\beta_0 + \beta_1 x_i) + (\varepsilon_i)$

The model can be split into two parts: the fixed or deterministic part, $(\beta_0 + \beta_1 x_i)$ and the random part, ε_i. Since ε_i is the random part, it has a distribution, which is a normal distribution with mean 0 and variance σ_ε^2. Without derivations here, we can obtain the probability distribution of Y from the probability distribution of ε.

Therefore, Y is normally distributed with mean $E(y_i) = \beta_0 + \beta_1 x_i$ and variances Var(Y) = σ_ε^2.The $E(y_i) = \beta_0 + \beta_1 x_i$ is the mean response or regression equation and is what will be estimated to obtain the regression line.

3.2.2 Estimation of the Regression Line

A line drawn on a scatter diagram that passes through majority of points and close to or divides other points into two equal or almost equal parts is called a regression line or line of best fit. Technically put, the line of best fit is the line which minimizes the sum of squares of the errors.

Regression equation is the equation of the line of best fit. It is a mathematical equation that enables the prediction of values of the dependent variable from known values of one or more independent variables. As stated earlier, much of the work done in regression analysis is to estimate the regression equation. The equation is given by $E(Y) = \beta_0 + \beta_1 X$

; where E(Y) is the mean response while β_0 and β_1 are the parameters to be estimated. The regression equation is often estimated by $\hat{Y} = b_0 + b_1 X$; where $\hat{Y}$ is the estimated mean response and gives the fitted values of Y, b_0 and b_1 are the point estimates of β_0 and β_1.

There three main methods of estimating or fitting the regression line: the eye-fitting method, the method of least squares and the method of maximum likelihood. For the purpose of this material the first two will be considered.

Eye-Fitting Method of Estimating the Regression Equation

After obtaining the line of best fit (the regression line), the linear equation of the line gives a subjective estimate of the regression equation.

Let the line equation be given by the general equation of a straight line,

$y = c + mx$, where m is the slope and c is the y-intercept. Comparing this to equation (2.4), we see that $b_0 = c$ and $b_1 = m$. This method is also referred to as observation method.

Let figure 2.1 below be a scatter plot of x_i, y_i bivariate data with a line of best fit on it.

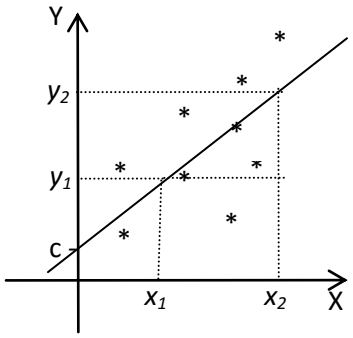


Fig. 2.1

Then $b_0 = c$ and $b_1 = \frac{y_2 - y_1}{x_2 - x_1}$.

Example 1:

The data below represents the number of units of a product (‘000 units) sold and the profit (₦’m) made. Using the method of Scatter diagram, determine the regression equation. Obtain the equation of the regression line.

Number of Units ('000) (X)	10	12	14	6	8	4
Profit (₦'m) (Y)	16	24	20	12	10	6

Solution:

The equation of the regression line is $\hat{y} = a + bx$, where *a* is the intercept and *b* is the slope.

In the scatter diagram below, the intercept *a* is obtained as the point where the fitted line cuts the y-axis

Note: it is difficult for the whole plotted points to fall on the fitted line but if they do, which is very rare, a perfect relationship exists between the two variables.

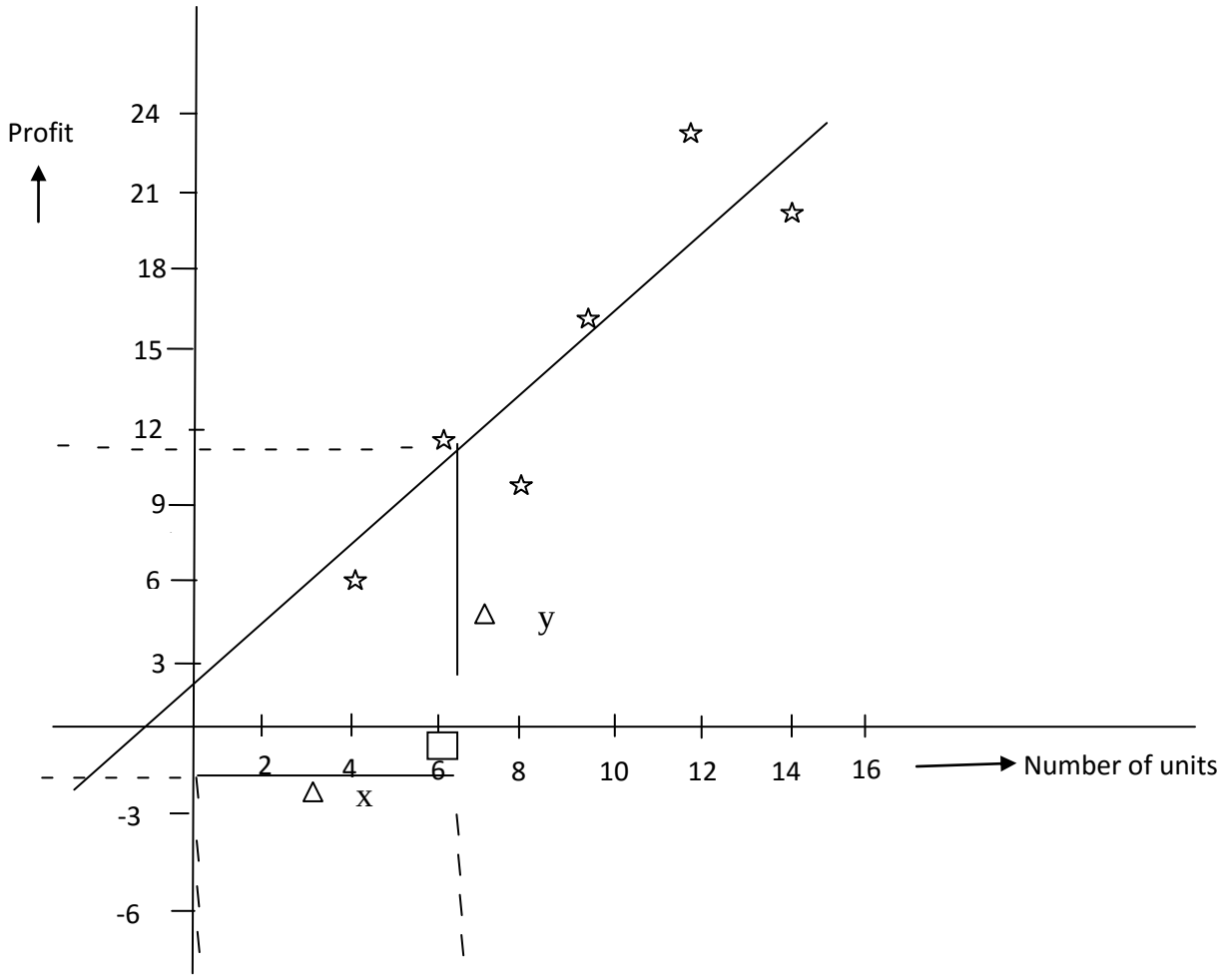
To estimate the value of *b*, a right angled-triangle of any size starting from the straight line and ending with the straight line so that the part of the straight line forms the hypotenuse of the triangle.

From the diagram below, one can obtain that

a= 2.8 and

$$b = \frac{\Delta y}{\Delta x} = \frac{16.5 - 8}{10 - 4} = \frac{8.5}{6} = 1.42$$

Therefore, $\hat{Y}$ = 2.8 + 1.42X



3.2.3 LEAST SQUARE ESTIMATION OF THE REGRESSION LINE

The regression line can often be fit by eye as discussed above, but, due to the differences in the line of best fit from different people, a more mathematical approach was considered to be the best method of fitting a regression line to bivariate data. The least squares formulas will be stated without derivations (the derivations if needed can be obtained from any standard statistical text).

The Least Square Estimators

The methods of least squares is employed in order to find good estimators for the regression parameters β_0 and β_1 . The method considers the deviation of Y_i from its expected value, $Y_i - (\beta_0 + \beta_1 x_i)$ for each (x_i, y_i).

The method requires that we consider the sum of n squared deviation given by:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

And minimizing Q gives the least squares estimators (formulas) for β_0 and β_1 as:

$$b_1 = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2},$$

And

$$b_0 = \tfrac{1}{n}(\sum y_i - b_1 \sum x_i) = \bar{y} - b_1 \bar{x}\,.$$

Where:

$\bar{x}$ is the mean value of the independent variable x, $\bar{y}$ is the mean value of the dependent variable y and b_0 and b_1 are called point estimators of β_0 and β_1, respectively.

Properties of Least Squares Estimators

When the assumptions of the regression model are met,

The least squares estimators b_0 and b_1 are unbiased; that is, $E(b_0) = \beta_0$ and $E(b_1) = \beta_1$

The least squares estimators have minimum variance among all unbiased linear estimators; that is, the sampling distributions of b_0 and b_1 have smaller variability than those of any other estimators. This also means that the least squares estimators are efficient.

The estimators will be such that the sum of the observed values Y_i equals the sum of the fitted values $\hat{Y}_1$. i.e, $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$.

The regression line always goes through the point $(\overline{X}, \overline{Y})$.

The Error Mean Square

The deviation of an observation Y_i from the estimated mean $\overline{Y}$ squaring it, and then summing over all such deviations is called sum of squares.

The error sum of squares, denoted by SSE is the sum of squares of Y from its expected value. It estimate, the residual sum of squares, is defined as:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 x_i\right)^2$$

The error sum of squares has n-2 degrees of freedom associated with it.

Therefore, the **error mean square**, denoted by MSE is given by:

$$MSE = \frac{SSE}{n-2} = \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{n-2}$$

$$= \frac{\sum \left(Y_i - b_0 - b_1 x_i\right)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

MSE is called error mean square or residual mean square and is an unbiased estimator of the error variance σ^2.

Example 2:

A farmer is interested in finding out if there is a relationship between the yield of his crop, maize and the application of fertilizer on his plots of land. He divided his farm into 11 plots and applied different concentrations of fertilizer. After harvest, the following results were obtained and tabulated.

Plot	1	2	3	4	5	6	7	8	9	10	11
Fert.	0	250	500	750	1000	1250	1500	1250	2000	2250	2500
Kg/acre	31	42	53	58	65	74	77	80	83	81	95

Estimate the regression coefficient and provide the regression equation.

Solution

Fertilizer (X)	Kg/acre (Y_i)	X_1Y_1	X^2
0	31	0	0
250	42	10500	62500
500	53	26500	250000
750	58	43500	562500
1000	65	56000	1000000
1250	74	92500	1502500
1500	80	115500	2250000
1750	83	140000	3062500
2000	81	166000	4000000
2250	18	182250	5062500
2500	95	237500	6250000
13750	745	1092750	24062500

For the data:

$n = 11, \quad \sum x = 13750, \quad \sum Y = 745, \quad \sum x^2 = 24062500$

$\sum XY = 1092750, \quad \bar{X} = 1250, \quad \bar{Y} = 67.73.$

To obtain b_1, we use.

$$b_1 = \frac{\sum X_1 Y_1 - \frac{\left(\sum X_1\right)\left(\sum Y_1\right)}{n}}{\sum X_1^{\,2} - \frac{\left(\sum X_1\right)^2}{n}} = \frac{1092750 - \frac{(13750)(745)}{11}}{24062500 - \frac{(13750)^2}{11}}$$

$b_1 = 0.0235$.

To obtain b_0, we use.

$$b_o = \frac{1}{n}\left(\sum Y_i - b_i \sum X_I\right) = \frac{1}{11}(745 - 0.235(13750)$$

$b_o = 38.36$

$$\therefore \hat{Y} = 38.36 + 0.0235x$$

3.3 INFERENCE CONCERNING REGRESSION PARAMETERS

One of the major uses of regression analysis is to make inferences for the future. In doing this, we make inference on the parameter in order to establish whether there is a linear associations between Y and X.

3.3.1 Inference concerning β_1 and β_1:

Inference concerning β_1

The sampling distribution of $b_1$is a normal distribution with mean $E(b_1) = \beta_1$ and variance $Var(b_1) = \sigma^2(b_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$; which can be estimated by $S^2(b_1) = \frac{MSE}{\sum(x_i - \bar{x})^2}$ = $\frac{MSE}{\sum x_i^2 - \frac{(\sum x_I)^2}{n}} = \frac{SSE/(n-2)}{\sum(x_i - \bar{x})^2}$, and the standard error of b_1 is $S(b_1) = \sqrt{Var(b_1)}$.

Therefore, the standardized statistic $b_1 - \beta_1 / \sigma(b_1)$is a standard normal variable and is distributed as $t_{(n-2)}$for the regression model. Hence, the $(1-\alpha)100\%$ confidence limits for β_1 are:

$$b_1 \pm t_{(\alpha/2, n-2)}\, S(b_1).$$

Test concerning β_1 are done using the t-distribution and may be in any of the forms given below. The most common is the two sided test described below. The hypotheses are:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

The Test statistic is

$$t_{cal} = \frac{b_1}{S(b_1)}$$

The decision rule with this test statistic at a given level of significance at α is:

Reject H_0 if $|t^*| > t_{(\alpha/2, n-2)}$, otherwise, do not reject H_0.

Inference concerning β_0:

When the scope of a model includes x = 0, inference about the intercept of the regression line might be required.

The point estimator of β_0 is given as:

$$b_0 = \bar{Y} - b_1 \bar{x}$$

The sampling distribution of b_0 is normal with

Mean, $E(b_0) = \beta_0$

and

Variance, $\sigma^2(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]$.

The normality of b_0 follows because, b_0, like b_1, is a linear combination of the observations y_i. An estimator of $\sigma^2(b_0)$ is obtained by replacing σ^2 by its point estimator MSE:

$$S^2(b_0) = MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_1 - \bar{x})^2}\right]$$

The square root, $S(b_0)$ is an estimator of $\sigma(b_0)$ and gives the standard error of b_0.

The statistic $(b_0 - \beta_0)/S(b_0)$ is distributed as $t_{(n-2)}$ for regression model.

Hence, $1-\alpha$ confidence interval for β_0 is

$(b_0 \pm t_{(\alpha/2, n-2)})S(b_0)$ and this is also set up in the same manner as β_1.

Test concerning β_0 is similar to that of β_1 by replacing b_1 by b_0.

Prediction

One of the major goals in regression analysis is prediction. The prediction of an outcome Y corresponding to a given level X of the independent variable is viewed as a result of a new trial on which the regression analysis is based. For any fitted model of the form $\hat{Y} = b_0 + b_1 X$, given any value of x or y, the corresponding value of the other variable can be obtained.

We denote the level of X for the new trial as X_n and the new observation on Y as Y_n, assuming that the underlying regression model applicable for the basic sample data continues to be appropriate for the new observation. Then we have, $\hat{Y}_n = b_0 + b_1 X_n$.

Example 3:

The relationship between the profit expected (Y, in ₦Million) and the number of units of the product sold (X in ₦'000) can be explain by the regression equation $\hat{Y}$ = 2.8 + 1.42X.

a.) Find the profit expected when the number of units of the product sold is 18,000

b.) Find the number of units sold if the profit made is ₦25 million.

Solution:

a.) the profit expected when the number of units of the product sold is 18,000can be obtained by

substituting 18 for x in model.

That is;

$\hat{Y}$ = 2.8 + 1.42 (18) = 28.36 =₦28.36m

b.) In this case, Y is given then we solve for X. Therefore, the number of units sold can be obtained substituting 25 for $\hat{Y}$ in the model.

That is, 25 = 2.8 + 1.42x and X = $\frac{25-2.8}{1.42}$ = 15.6338 = 15, 634 units

Example 4:

A farmer is interested in finding out if there is a relationship between the yield of his crop, maize and the application of fertilizer on his plots of land. He divided his farm into 11 plots and applied different concentrations of fertilizer. After harvest, the following results were obtained and tabulated.

Plot	1	2	3	4	5	6	7	8	9	10	11
Fert.	0	250	500	750	1000	1250	1500	1250	2000	2250	2500
Kg/acre	31	42	53	58	65	74	77	80	83	81	95

a) Estimate the regression coefficient and provide the regression equation.

b) What would be the yield when fertilizer concentration of 3000 is applied on a new plot.

Solution:

Fertilizer (X) Kg/acre	(Y_i)	X_1Y_1	X^2
0	31	0	0
250	42	10500	62500
500	53	26500	250000
750	58	43500	562500
1000	65	56000	1000000
1250	74	92500	1502500
1500	80	115500	2250000
1750	83	140000	3062500
2000	81	166000	4000000
2250	18	182250	5062500
2500	95	237500	6250000
13750	745	1092750	24062500

a) From the data:

$$n = 11, \quad \sum x = 13750, \quad \sum Y = 745, \quad \sum x^2 = 24062500$$
$$\sum XY = 1092750, \quad \bar{X} = 1250, \quad \bar{Y} = 67.73.$$

To obtain b_1, we use.

$$b_1 = \frac{\sum X_1Y_1 - \frac{\left(\sum X_1\right)\left(\sum Y_1\right)}{n}}{\sum X_1^{\,2} - \frac{\left(\sum X_1\right)^2}{n}} = \frac{1092750 - \frac{(13750)(745)}{11}}{24062500 - \frac{(13750)^2}{11}}$$

$b_1 = 0.0235.$

To obtain b_0, we use.

$$b_o = \frac{1}{n}\left(\sum Y_i - b_i \sum X_I\right) = \frac{1}{11}(745 - 0.235(13750)$$

$b_o = 38.36$

$$\therefore \hat{Y} = 38.36 + 0.0235x$$

b) When x = 3000, $\hat{Y} = 38.36 + 0.0235(3000) = 38.36 + 70.5 = 108.86$,

3.3.2 Testing for Overall Significance of the Regression Model

The test involves first partitioning the sums of squares and degrees of freedom associated with the response variable Y.

Sum of Squares:

The sum of squares is a measure of variability. The total variability due to the response variable Y is usually called the total sum of squares. The total sum of squares can be split into two, sum of squares due to regression of Y on X and sum of squares due to random error.

We define the following terms:

a. **Total Sum squares (SST):**

SST is the measure of total variation. It is the sum of the squared deviations $\left(Y_i - \overline{Y}\right)^2$.

$\therefore SST = \sum \left(Y_i - \overline{Y}\right)^2$

If SST = 0, all the observations are the same.

b. **SSE (Error Sum of squares):**

This is the measure of variation in the data with the regression model.

$SSE = \sum \left(Y_i - \hat{Y}\right)^2$

If SSE = 0, all observations fall on the fitted regression line.

c. **SSR (Regression Sum of Squares)**.

SSR is the sum of the squared deviations, $\hat{Y}_i - \overline{Y}$. It is the difference between the fitted values on the regression line and the mean of the observed values.

Such that,

$$\sum\left(Y_i - \bar{Y}\right)^2 = \sum\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum\left(Y_i - \hat{Y}_i\right)^2$$

$\therefore$ SST = SSR +SSE

Degrees of Freedom

If n is the number of observations and p the number of parameters, then the degrees of freedom for the total sum of squares, regression sum of squares and the error sum of squares are (n – 1), (p - 1) and (n - p) respectively.

Therefore, for the simple linear regression model with only two parameters, the degree of freedom due to regression sum of squares (SSR) is 1, the degree of freedom due to error sum of squares (SSE) is (n – 2), while the total degree of freedom is (n - 1).

Notice that as the regression and error sums of squares partitions the total sum of squares, so are their degrees of freedom. That is,

(n - 1) = (n - p) + (p - 1).

Mean Squares

A sum of squares divided by its associated degrees of freedom is called mean square (MS). For simple linear regression model,

The regression mean square (MSR) is given by:

$$\text{MSR} = \frac{SSR}{1} = SSR.$$

The error mean square (MSE) is given by:

$$MSE = \frac{SSE}{n-2}$$

Test of Hypothesis for Overall Regression

We may wish to test the following hypothesis.

H_0: There is no significant regression

H_1: There is a significant regression

Comments:

The mean squares are actually variances.

MSR follows a chi-square distribution with 1 degree of freedom.

MSE follows a chi-square distribution with (n – 2) degrees of freedom

The ratio of two chi-squares follows an F-distribution with numerator and denominator degrees of freedom as it parameters.

If H_0 holds, the test statistic for the test of overall regression is given by: $F_{cal} = \frac{MSR}{MSE} \sim F_{(a,b)\alpha}$; where a is regression degree of freedom (numerator) and b is the error degree of freedom (denominator).

ANOVA table for Simple Linear Regression

Source of Variation	SS	df	MS	F*
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
Error	$SSR = \sum(Y_i - \hat{Y}_i)^2$	n-2	$MSE = SSE/n-2$	
Total	$SST = \sum(Y_i - \bar{Y}_i)^2$	n-1		

Since the test is upper-tailed and F* is distributed as $F_{\alpha(1,n-2)}$ when H_o holds, the decision rule, when the risk of type 1 error is to be controlled at α, is as follows

If $F^* > F_{(\alpha\ ;\ 1,\ n-2)}$, reject H_a

Where $F_{\alpha;\ (1,n-2)} = F_{crit}$ is the critical value of the F distribution.

3.3.3 Coefficient of Determination and Standard Error of Estimate

Coefficient of Determination

The coefficient of determination, denoted by r^2, is a measure of the effect of X in reducing the variation in Y. It gives the proportionate reduction of the total variation associated with the use of the independent variable X, that is, it is the proportion of the variability in Ydue to X. It is used to answer the research question: To what extent does X affect Y?

We have

$$r^2 = \frac{SST - SSE}{SST} = \frac{SSR}{SST}$$

And, $0 \leq r^2 \leq 1$.

The following should also be noted about "r"

i. If SSE = 0, $r^2 = 1$. Here, X accounts for all variation in Y.

ii. If $b_1 = 0$, so that $\hat{Y}_i - \overline{Y}$,SST – SSE and $r^2 = 0$. Then there is no linear association between X and Y in the sample data.

A value of r or r^2 relatively close to 1, sometimes is taken as an indication that sufficiently precise inferences on Y can be made from the knowledge of X.

Standard Error of Estimate (S_e)

Standard error of estimate (S_e) is a measure developed by statisticians to assess the reliability of the fitted regression line, hence it measures the variability or scattering of the observed values of 'y' around the regression line. The larger the 'S_e', the more significant are the magnitude of the individual variations and less reliable the regression line. Therefore, a smaller S_e is preferred.

Standard error of estimate can be evaluated using the expression:

$$S_e = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$$

Example 4:

A farmer is interested in finding out if there is a relationship between the yield of his crop, maize and the application of fertilizer on his plots of land. He divided his farm into 11 plots and applied different concentrations of fertilizer. After harvest, the following results were obtained and tabulated.

Plot	1	2	3	4	5	6	7	8	9	10	11
Fert. (X)	0	250	500	750	1000	1250	1500	1250	2000	2250	2500
Yield (Y)	31	42	53	58	65	74	77	80	83	81	95

a.) Estimate the regression coefficient and provide the regression equation.

b.) Test the hypothesis which states that the regression of maize yield on fertilizer application is not significant putting $\alpha = 0.05$.

c.) What percentage of the maize yield is due to the concentration of fertilizer applied?

d.) Compute the standard error of estimate.

Solution

Fertilizer (X)	**Yield (Y_i)**	**X_1Y_1**	**X^2**	**Y^2**
0	31	0	0	961
250	42	10500	62500	1764
500	53	26500	250000	2809
750	58	43500	562500	3364

1000	65	65000	1000000	4225
1250	74	92500	1562500	5476
1500	77	115500	2250000	5929
1750	80	140000	3062500	6400
2000	83	166000	4000000	6889
2250	81	182250	5062500	6561
2500	95	237500	6250000	9025
13750	**739**	**1079250**	**24062500**	**53403**

a) For the data:

$n = 11, \quad \sum x = 13750, \quad \sum Y = 739, \quad \sum x^2 = 24062500$

$\sum XY = 1079250, \quad \bar{X} = 1250, \quad \bar{Y} = 67.18$

To obtain b_1, we use.

$$b_1 = \frac{\sum X_1 Y_1 - \frac{\left(\sum X_1\right)\left(\sum Y_1\right)}{n}}{\sum X_1{}^2 - \frac{\left(\sum X_1\right)^2}{n}} = \frac{1079250 - \frac{(13750)(739)}{11}}{24062500 - \frac{(13750)^2}{11}} = \frac{1079250 - 923750}{24062500 - 17187500} = \frac{155500}{6875000} = 0.0226$$

To obtain b_0, we use.

$b_o = \bar{Y} - b_1 \bar{X} = 67.18 - 0.0226(12500) = 38.93$

$\therefore \hat{Y} = 38.93 + 0.0226X$

This means that for every unit increase in the quantity of fertilizer, there is an increaseof 0.0226 in the yield of maize.

b. we use the ANOVA test for simple regression the hypotheses are

H_0: The regression is not significant

H_1: The regression is significant

Following from the table above, we have:

Yhat	(Y-Yhat)sq	(Yhat-Ybar)sq	(Y-Ybar)sq
38.93	62.8849	798.0625	1308.992
44.58	6.6564	510.76	634.0324
50.23	7.6729	287.3025	201.0724
55.88	4.4944	127.69	84.2724
61.53	12.0409	31.9225	4.7524
67.18	46.5124	2.02E-28	46.5124
72.83	17.3889	31.9225	96.4324
78.48	2.3104	127.69	164.3524
84.13	1.2769	287.3025	250.2724
89.78	77.0884	510.76	190.9924
95.43	0.1849	798.0625	773.9524
738.98	**238.511**	**3511.475**	**3755.64**

Hence,

$$SST = \sum \left(Y_i - \bar{Y}\right)^2 = 3755.64$$
$$SSR = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 = 3511.475$$
$$SSE = \sum \left(Y_i - \hat{Y}_i\right)^2 = 238.511$$

The ANOVA Table:

Source	SS	Df	MS	F*
Regression	3511.475	1	3511.503	132.49
Error	238.511	9	26.501	

Total	3755.64	10		

Since $\alpha = 0.05,$ and $n = 10;\ F_{(.05;\,1,\,9)} = 5.12.$

Since calculated F value (132.49)> tabulated $F_{(.05,1,9)}$, H_0 is rejected.

That is, the regression of maize yield on fertilizer is significant.

c) From the ANOVA table,

$SST = 3755.64$

$SSR = 3511.503$

$$\therefore\ r^2 = \frac{SSR}{SST} = \frac{3511.475}{3755.64} = 0.935$$

This implies that 93.5% of of the maize yield is due to the concentration of fertilizer.

d.) The standard error of estimate is:

$$S_e = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$$

From the table above, we have:

$\Sigma Y^2 = 53403$, $\Sigma Y = 739$, $\Sigma XY = 1079250$, $a = 38.93$, $b = 0.0226$

$$\therefore S_e = \sqrt{\frac{53403 - (38.93)(739) - (0.0226)(1079250)}{11-2}} = \sqrt{\frac{53403 - 28769.27 - 24391.05}{9}},$$

$$= \sqrt{\frac{242.68}{9}} = \sqrt{\frac{242.68}{9}} = \sqrt{26.9644} = 5.193.$$

3.4 MULTIPLE LINEAR REGRESSION

Multiple regression can be considered to be any regression problem with the number of parameters, p > 2. In other words, the number of independent variables is more than one.

3.4.1 The Multiple Linear Regression Model

The general linear form of this regression model, with normal error terms is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i(.p-1)} + \varepsilon_i$$

where:

y_i = observations

i = 1, 2, …, n

$\beta_0, \beta_1, ..., \beta_{p-1}$ are parameters

$x_{i1}, x_{i2}, ..., x_{i.p-1}$ are independent or explanatory variables

ε_i are error terms, which are independently and identically normally distributed with mean 0 and variance σ^2.

The mean response function for the multiple regression model is,

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1}; \text{ since } E(\varepsilon_i) = 0.$$

The parameters $\beta_1, \beta_2, ..., \beta_{P-1}$ are frequently called partial slope because they reflect the partial effect of one independent variable when the other independent variables are included in the model and is held constant while β_0 is the y intercept of the regression plane.

The general Linear Regression Model in Matrix form

To express the general linear model:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_{p-1} x_{1.p-1} + \varepsilon_i$$

in matrix terms, we define the following matrices:

$$\underset{nX1}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \text{ with } Y' = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}, \text{ where } Y' \text{ is the transpose of Y.}$$

$$\underset{nXp}{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(p-1)} \end{pmatrix}$$

The matrix X consists of a column of 1's and p-1 columns containing the n-values of the p-1 independent variables. Although normally referred to as design matrix, X is a matrix of coefficients of the parameters.

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Then the regression model in matrix form is given by:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(p-1)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Where,

$\boldsymbol{Y}$ is a vector of observations on the dependent variable

B is a vector of parameters

X is a matrix of constant coefficients or design matrix

$\boldsymbol{\varepsilon}$ is a vector of independent normal random variable with expectation E(ε) = 0 and variance, $\sigma_{\varepsilon}^{2} = \sigma^{2} I$

E($\boldsymbol{Y}$) = $\boldsymbol{X\beta}$ is a vector of expected values for the model and the varianceof Y is $\sigma^{2}\boldsymbol{I}$.

Note that the general regression model also include simple linear regression model, when the number of parameters p = 2.

3.4.2 Least Squares Estimators of the Parameters

Estimation of the general regression model can also be done using matrix approach. However, computer packages have made the implementations easy.

Given the general regression model, the least squares normal equations are:

$$\underset{p\times p}{X'X}\ \underset{p\times 1}{b} = \underset{p\times 1}{X'Y},$$

Which can be solved to give

$$\therefore b = (X'X)^{-1}(X'Y);$$

Where,

$b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{pmatrix}$, is the estimate of the parameter vector $\boldsymbol{\beta}$.

These least squares estimators are unbiased, efficient, consistent and have minimum variance.

Note that when the number of independent variables is 2, the following can be used.

$$X'X = \begin{bmatrix} n & \sum x_{11} & \sum x_{12} \\ \sum x_{11} & \sum x_{11}^2 & \sum x_{11}x_{12} \\ \sum x_{12} & \sum x_{12}x_{11} & \sum x_{12}^2 \end{bmatrix} \text{ and } \quad X'Y = \begin{bmatrix} \sum y_i \\ \sum x_1 y_i \\ \sum x_{12} y_i \end{bmatrix}.$$

For the general regression model, computer packages have made the implementations easy for both estimation and inference.

Example 5:

Daystar Hospital Plc operates in 21 cities of medium size. These hospitals specialize in surgery of children more than adults. The hospital is considering an expansion into other cities of medium size and wishes to investigate whether profit (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1) and the per capita disposable personal income in a community (X_2). Data on these variables for the most recent year for 21 cities in which Daystar hospital is now operating are shown below. Profits are expressed in thousands of dollars and are labeled Y; the number of persons aged 16 or younger is expressed in thousands of persons and labeled X_1 and per capita disposable personal income is expressed in thousands of dollars and labeled X_2.

If the multiple regression model: $y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \varepsilon_i$ is appropriate, obtain the estimated regression function.

i	1	2	3	4	5	6	7	8	9	10	11
Y	174.4	164.4	244.2	154.6	181.6	207.5	152.8	163.2	145.4	137.2	241.9
X_1	68.5	45.2	91.3	47.8	46.9	66.1	49.5	52.0	48.9	38.4	87.9
X_2	16.7	16.8	18.2	16.3	17.3	18.2	15.9	17.2	16.6	16.0	18.3

I	12	13	14	15	16	17	18	19	20	21
Y	191.1	232.0	145.3	161.1	209.7	146.4	144.0	32.6	224.1	165.5
X_1	72.8	88.4	42.9	52.5	85.7	41.3	51.7	89.6	82.7	52.3
X_2	17.1	17.4	15.8	17.8	18.4	16.5	16.3	18.1	19.1	16.0

Solution:

The X and Ymatrices for the Daystar hospital is as follows:

$$X = \begin{bmatrix} 1 & 68.5 & 16.7 \\ 1 & 45.2 & 16.8 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 52.3 & 16.0 \end{bmatrix}, \quad Y = \begin{bmatrix} 174.4 \\ 164.4 \\ . \\ . \\ . \\ 166.5 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ 68.5 & 45.2 & . & . & . & 52.3 \\ 16.7 & 16.8 & . & . & . & 16.0 \end{bmatrix} \begin{bmatrix} 1 & 68.5 & 16.7 \\ 1 & 45.2 & 16.8 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 52.3 & 16.0 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 21.0 & 1{,}302.4 & 360.0 \\ 1302.4 & 87{,}707.9 & 22{,}609.2 \\ 360.0 & 22609.2 & 6190.3 \end{bmatrix}$$

Also,

$$X'Y = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ 68.5 & 45.2 & . & . & . & 52.3 \\ 16.7 & 16.8 & . & . & . & 16.0 \end{bmatrix} \begin{bmatrix} 174.4 \\ 164.4 \\ . \\ . \\ . \\ 166.5 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 3820 \\ 249643 \\ 66073 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 21.0 & 1302.4 & 360.0 \\ 1302.4 & 87707.9 & 22609.2 \\ 360.0 & 22609.2 & 6,190.3 \end{bmatrix}^{-1}$$

$$(X'X^{-1}) = \begin{bmatrix} 29.7289 & .0722 & -1.9926 \\ .0722 & .00037 & -.0056 \\ -1.9926 & -.0056 & .1363 \end{bmatrix}$$

$$\therefore \quad b = (X'X)^{-1} X'Y = \begin{bmatrix} 29.7289 & .0722 & -1.9926 \\ .072 & .00037 & -.0056 \\ -1.9926 & -.0056 & .1363 \end{bmatrix} \begin{bmatrix} 3820 \\ 249,643 \\ 66073 \end{bmatrix}$$

$$\therefore \quad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix}$$

and the estimated regression function is

$$\hat{y} = -68.857 + 1.455\,x_1 + 9.366x_2 .$$

Interpretation:

This estimated regression function indicates that mean sales are expected to increase by 1.455 thousand dollars when the target population increases by 1 thousand persons aged

16 years or younger, holding per capita personal income constant, and that mean sales are expected to increase by 9.366 thousand dollars when per capita income increase by 1 thousand dollars, holding the target population constant.

4.0 CONCLUSION

In this unit, correlation, simple linear regression and multiple linear regression have been considered. Definitions, their uses in analysis as well as hypothesis testing have been presented.

5.0 SUMMARY

In this module, the application of regression and correlation analyses have been demonstrated with simple illustrative practical examples that are capable of guiding any user to appreciate the regression and correlation tools to take valuable decisions.

6.0 TUTOR-MARKED ASSIGNMENT

1.a) Distinguish between Correlation and Regression.

b) The quantities of rice (x) and beans (y) purchased by a group of friends for annually celebration is tabulated below:

Year	1981	1982	1983	1984	1985	1986	1987
Rice (x)	118	112	101	98	94	91	62
Bean (y)	76	57	58	80	70	43	53

Compute and compare the Pearson moment correlation and the spearman rank correlation and comment on your result.

2. The marks obtained by ten students in Physics and Mathematics are given below:

Physics (p)	79	63	84	46	77	73	56	58	49	69

Mathematics (m)	56	42	59	35	54	62	47	51	24	49

Make a scatter plot of physics and Maths and comment on the degree of association between them.

Calculate the Spearman correlation coefficient and comment on your result.

Compute the Pearson's moment correlation coefficient.

Test the hypothesis H_0: $\rho = 0$ against H_1: $\rho \neq 0$ at 5% level of significant and draw appropriate conclusion.

3a. Define regression.

b. Distinguish between simple linear regression and multiple linear regression.

4. The table below gives the distribution of the population in 8 small cities over a period of 80 years.

Year (y)	10	20	30	40	50	60	70	80
Population p(000)	9.0	10.2	12.0	13.9	15.9	17.9	20.1	22.9

If the population p, is linearly related to the year y, that is $p = n + ky$, where n and k are constants.

Plot the scatter diagram

Draw an eye fitted line of best fit

Use your graph to estimate the values of n and k and then population in the year 90.

5. Given the data below:

Physics (p)	79	63	84	46	77	73	56	58	49	69
Mathematics (m)	56	42	59	35	54	62	47	51	24	49

What will be the score of a student in physics. If he scores 61in Maths.

Fit a suitable regression line to the information above.

6a). The model of a simple linear regression is $Y_i = \beta_o + \beta_1 X_i + \varepsilon_i$

State what $Y_i, \beta_0, \beta_1, X_i \text{ and } \varepsilon_i \text{ represent.}$

b.) When asked to state the simple linear regression model, a student wrote it as: $E(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$ Do you agree?

7. A company is to take delivery of 20 incoming shipments of chemicals in drums arriving at their warehouse, x_1 is the number of drums in shipment, x_2 (in hundred pounds), is the total weight of shipment, and y represents the number of minutes required to handle shipment. The data are given below:

I	1	2	3	4	5	6	7	8	9	10
x_2	7	18	5	14	11	5	23	9	16	5
x_1	511	16.72	3.20	7.03	10.98	4.04	2.07	7.03	10.62	4.76
y_i	58	152	41	93	101	38	203	78	117	44

I	11	12	13	14	15	16	17	18	19	20
x_2	17	12	6	12	8	15	17	21	6	11
x_1	11.02	9.51	3.79	6.45	4.60	13.86	13.03	15.1	3.64	9.57
y_i	121	112	50	82	48	127	140	155	39	90

(a) obtain the estimated regression function. How is b_1 here interpreted? How is b_2 here interpreted

(b) Test whether there is a regression relation, using a level of significance of 0.05. State the alternatives, decision rule, and conclusion.

(c) What does your test result imply about β_1 and β_2?

7.0 REFERENCES/FURTHER READING

i. Onyeka-Ubaka, J. N. (2013). *Multi-Level Statistics: An Academic Companion for Inter-Disciplinary Professional Competence*, Royal Choice Multi-Media, Lagos.

ii. Osuagwu, L. (2002). Business Research Methods: Principles and Practice. Lagos: Gre Resource Ltd.

iii. Spiegel, A. (2011). Statistics. New York, Schum Series.

iv. Rose, B. (1995). Fundamentals of Biostatistics (4^{th} ed.), Duxbury Press, ITP, USA.

v. Le, C. T. (2003). Introductory Biostatistics. John Wiley and Sons Publishing, Wiley-Interscience.

vi. Department of Mathematics (2015). A First Course in Statistics. University of Lagos, Lagos: Nile Ventures.