COURSE GUIDE

PHS 210
INTRODUCTION TO BIOSTATISTICS

Course Team Professor Chikaike Ogbonna (Course Developer/Writer) - UNIJOS
Dr. Tolulope Afolaranmi (Course Developer/Writer) - UNIJOS
Dr. Gloria Anetor (Course Coordinator) – NOUN
Dr. Jane Agbu (Programme Leader) - NOUN

NATIONAL OPEN UNIVERSITY OF NIGERIA

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

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

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

Published by
National Open University of Nigeria

Printed 2018

ISBN: 978-978-8521-86-0

CONTENTS **PAGE**

INTRODUCTION

This course, PHS 210 is a three-credit unit course. Biostatistics is a biological science and a branch of medicine that involves the processes of collection, collation, analysis, interpretation, presentation of health and health related information that may lead to conclusion and decision making. Biostatistics therefore embraces those techniques pertaining to health and health related sciences. This is further exemplified in that although the methodology in statistical processes are general, especially, in basic principles and concepts such are best appreciated by medically orientated persons if placed in a familiar context. This is more so because specific problems arise in health sciences where the unit of interest is the **living person** and not some sort of abstract phenomena, objects, items or finance. Biostatistics involves deductive and inductive inferences. In deductive application, we reason from the general to the specifics while in the inductive we observe the specifics such as individuals or particulates to infer on the general. Inductive biostatistics is scientifically inclined and follows probabilistic theory and therefore usually leads to scientific conclusion. This course guide tells you what to expect from reading this course material.

WHAT YOU WILL LEARN IN THIS COURSE

The study of Biostatistics will equip you with the knowledge, skills and competences in generating health and health related data. It will educate you on appropriate choice of test statistics and methods of statistical analysis. You will be able to acquire statistical skills for data analysis and contextual interpretation of statistical results effectively and with less difficulty.

COURSE AIM

The aim of this course is to provide you with the necessary fundamental knowledge, skills and understanding of Biostatistics and its application in health and health-related fields in the principles and practice of Public Health.

COURSE OBJECTIVES

After going through this course, you should be able to:

- define, explain the scope and concepts of Biostatistics and its applications.
- define data, know the common sources of data and understand the various methods of data collection and presentation.

- discuss to a reasonable level the probability theories and population distribution and their applications in health and health related fields.
- define and explain the various commonly used test statistics and their applications in analyzing data in public health
- acquire the necessary knowledge and skills for data analysis and interpretation of test statistic results.

WORKING THROUGH THIS COURSE

This course has been carefully packaged to bridge the gap between two extremes of complex statistical and elementary methods and at the same time provide the basic essentials in Biostatistics for public health scholars. You will therefore notice that efforts have been made to simplify complex calculations with relevant worked examples and illustrations to enhance your interest in understanding statistical applications. You should take advantage of the user friendliness of this Biostatistics course to acquire the knowledge and skills in Biostatistics' principles and methods. You are therefore expected to be studious and also ensure regular attendance in tutorial sessions, participating in group discussions and not hesitant to ask questions where necessary.

COURSE MATERIALS

This course comprises of 10 modules broken into 28 units. They are as listed below.

i. A Course guide
ii. Study Units

STUDY UNITS

This course comprises of ten modules broken down into 28 units. They are as listed below.

Module 1 Introduction to Biostatistics and Data Management

Unit 1 Definition and Application of Biostatistics
Unit 2 Data, Data Sources
Unit 3 Methods of Data Collection
Unit 4 Measuring Instruments

Module 2 Screening Tests and Variables

Unit 1 Defining Screening Tests
Unit 2 Variables and Classification of Variables

Module 1

In Unit 1, you will be taught the meaning and the application of Biostatistics in public health. The unit will also treat the various nomenclatures and applications in Biostatistics. In Unit 2, you will understand the meaning of data, various sources of data and in Unit 3 the different methods of Data Collection in public health. In Unit 4, you will be taught the various types of measuring instruments used in data generation in health and health- related fields.

Module 2

In Unit 1, you will know the meaning and components of Screening Tests and their applications in the practice of public health. In Unit 2, variable will be defined and explained. Also, the various classifications used in describing variables will be discussed.

Module 3

In Unit 1, data presentation using various types of tables will be discussed in details. The qualities of a good tabular presentation and appropriate choice of tables for various types of variables will also be treated. In Unit 2 diagrammatic presentation of data will be discussed in details. Various types of diagrams will be explained and the appropriate diagram for the right variables will be taught and demonstrated. You will be able to choose the right diagrams for various data generated in order to clearly and appropriately describe the spread of the population characteristics visually. In Unit 3, the presentation of data using Maps and various graphs will be discussed.

Module 4

In Unit 1, the various measures of central tendency also known as averages i.e. Mean, Median and Mode will be discussed in details and their place and application in public health explained. In Unit 2, various measures of location will be mentioned and discussed including their application. In Unit 3, the measures of dispersion, spread or variability will be treated with emphasis on the commonly used ones namely standard deviation, variance and coefficient of variation. How to calculate them and their uses, place and application in public health will be explained.

Module 5

In Unit 1, measures of relationship such as correlation and regression will be discussed. How to calculate them and their applications will also

be explained. In Unit 2 you will learn about the probability and probability theories including its applications in Biological sciences. There will be examples of natural and genetic processes that following probability theories, laws and rules in life.

Module 6

In Unit 1, Normal distribution or the Gaussian curve and the features of the distribution will be discussed. In Unit 2 Standard Normal distribution will be discussed and the features listed and explained. In Unit 3, the Binomial and Poisson distribution theories will be discussed with worked examples for clearer understanding of the principles and applications. In Unit 4, Skewed distribution will be discussed and various models explained. Their applications and graphic presentation will be provided for better understanding of their application.

Module 7

In Unit 1, Sampling in the context of Biostatistics will be defined and explained. Definition of terms used in sampling will be provided. The meaning, concept and various Samples will be explained. In Unit 2, you will be taught in details the meaning of Non-Probability Sampling Techniques, the various types of non-probability sampling techniques and when they are applicable in public health practice. In Unit 3,the meaning and concept of Probability Sampling Technique will be discussed in details. The various types of probability sampling techniques will be explained and you will be taught the appropriate application of each sampling technique.

Module 8

In Unit 1, Inferential Biostatistics will be defined and the concepts in use will be explained while the classification based on the concepts discussed. In Unit 2, you will learn about the meaning of hypothetical statement, the concept and types of hypotheses. Hypothesis testing and the steps in the process of hypothesis testing will be discussed. In Unit 3 Test of significance will be defined and discussed. The application of test statistics will be explained. You will learn about the formula and calculation of various test statistics. You will also learn how to choose the appropriate test statistic and significant levels and understand statistical Errors.

Module 9

In Unit 1, the concept, definition and application of Non-parametric Tests will be discussed using Chi-Square Test as an example. The Chi-

Square test will be defined and the application discussed. You will be taught how to manually and with the use of appropriate software to analyse using Chi-Square test Statistics. There will be worked examples for better understanding. In Unit 2 other non-parametric tests will be listed and explained with the explanation on how to use them in non-parametric data analysis.

Module 10

In Unit 1, we will define and explain the concepts in parametric tests and the applications discussed. Parametric tests will be listed and how to use them for appropriate parametric data calculations discussed. How and when to use T-test will be discussed with worked examples. In Unit 2, Analysis of Variance (ANOVA) and Analysis of Co-variance (ANCOVA) will be defined and discussed. Their applications including how and when to use them will be explained.

TEXT BOOKS AND REFERENCES

The followings are list of textbooks, journals and website addresses that can be consulted for further reading:

Angela, Hebel. (2002). Parametric versus nonparametric statistics when to use them and which is more powerful. Psyphz.psych.wisc.edu/-sha-vs-nonparametricstats.ppt.

Bartlett, J. E., Kotrlik, J. W and Higgins, C. C. (2001). "Organizational Research: Determining Appropriate Sample Size in Survey Research."*Information Technology, Learning and Performance Journal*. *19* (1) 43- 50.

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York. 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York. 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Dorling Kindersley Pvt. Ltd. Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications, Ltd. Malden: 3-15.

Kothari, C. R. (2013). *Research methodology, Methods and Techniques*. 2nd Revised edn. New Age International (P) Ltd. New Delhi. 1-21.

Louis, J. C., Andres, de Francisco., Thomas, Nchinda et al. (2001-2002). The 10/90 Report on Health Research. Global Forum for Health Research. 12.

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers.*7th edition. Jaypee Brothers Medical Publishers (P) Ltd. New Delhi. 33-52.

Merida, L. J. (1997). *Information Management for Health Professions*. Delmar Publishers. Delmar. 50-71.

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi. 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Osuala, E. C. (2005). *Introduction to Research Methodology*. Third Edition. Africana-First Publishers Limited. Onitsha: 28-44.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Edinburgh: Blackwell Scientific Publications Ltd. 40-46.

Richard, F. Morton and Richard, J. Hebel. (1979). *A study guide to Epidemiology and Biostatistics*. Baltimore, New York: University Press.. 59-109.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Kano: Habason Nig. Limited.. 32-34.

Staffordshire University. Harvard Referencing System. Guide and Examples. www.staffs.ac.uk/support_depts/infoservices/learning_support/refzone/Harvard/.

Stover, John (1998). "Revising the Proximate Determinants of Fertility Framework: What Have We Learned in the past 20 Years?" *Studies in Family Planning. 29* (3): 255ï 267.

Syvia, Wassertheil-Smoller. (1990). *Biostatistics and Epidemiology. A primer for health professionals.* New York: Springer-Verlag. 119.

Wayne, W. D. (2006). *Biostatistics. A Foundation for Analysis in the Health Sciences*. 7th edition. John Wiley and Sons. New Delhi: 57-71.

Some Websites

- *www.bestdoctors.com*.
- www.healthatoz.com.
- www.healthcommunities.com.
- www.healthfinder.gov.
- www.intelihealth.com.
- PubMed/MEDLINE https://www.nlm.nih.gov>pmresources
- UMLS https://www.nlm.nih.gov>research>umls
- MedlinePlus https://medline.gov/
- TOXNEThttps://toxnet.nlm.nih.gov/
- LocatorPlushttps://locatorplus.gov>Pwebrecon
- NLMDatabaseshttps://wwwcf.nlm.nih.gov>eresources
- APIS Databases https://www.programmableweb.com

ASSESSMENT

There are two components of assessment for this course namely the tutor-marked assignments and the final examination.

TUTOR-MARKED ASSIGNMENT

The Tutor-Marked Assignment (TMA) is the continuous assessment component of your course. It accounts for 30 per cent of the total course. The TMAs will be submitted to your facilitator after you have done the assignment.

FINAL EXAMINATION AND GRADING

The examination concludes the assessment for the course. It constitutes 70 per cent of the whole course. You will be informed of the time for the examination.

SUMMARY

This course covers both Descriptive and Inferential Biostatistics and provides you with the relevant basic knowledge and skills in various methods of Biostatistics. It is hoped that you will be able to apply confidently and effectively too the acquired knowledge and competencies in the principles and practice public health. We wish you success in taking this course.

MAIN COURSE

CONTENTS **PAGE**

MODULE 1 INTRODUCTION TO BIOSTATISTICS AND DATA MANAGEMENT

UNIT 1 DEFINITION AND APPLICATION OF BIOSTATISTICS

CONTENTS

1.0 INTRODUCTION

Through our God given senses we observe and experience the world around us. Our observation and experience stimulate scientific reasoning. Therefore, we need to draw logical conclusion about the truth of our observation and experience. This course deals with statistical methods and data analysis in health and health-related fields, including the planning of data collection in terms of the design of surveys and experiments. It has treated the basic concepts and applications in Biostatistics that is enough to provide you with adequate knowledge and competencies required to generate and effectively analyze data and contextually interpret statistical results with little or no difficulty.

2.0 OBJECTIVES

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

- define and classify Biostatistics
- define and explain the concepts in Biostatistics
- explain the principles in the application of Biostatistics in public health.

3.0 MAIN CONTENT

3.1 Definition of Biostatistics

Biostatistics is the scientific process of collecting, collating, analyzing, presenting, interpreting and disseminating health and health-related data in estimating the magnitude of associations and testing hypotheses. Statistics therefore are pieces of facts or information generated from items, individuals or numbers that may form values such as averages and rates. Biostatistics can further be defined as a biological science and a branch of medicine that involves the processes of collection, collation, analysis, interpretation, presentation of health and health- related information that may lead to conclusion and decision-making.

The word statistics is derived from the Latin word **"state"** indicating the historical importance of governmental data gathering, which related principally to demographic information including census data and vital statistics and often to their use in military recruitment and tax collection. As biological entities are counted or measured, it becomes apparent that some objective (statistical) methods are necessary to aid the investigator in presenting and analysing such research data. Hence, the name Biostatistics - the application of statistical principles to biological measurements.

Biostatistics is all about ćhe Logic of Scientific reasoningô in uncovering the illusive truth. Biostatistics therefore embraces those techniques pertaining to health and health-related sciences. This is further exemplified in that although the methodology in statistical processes are general, especially, in basic principles and concepts such are best appreciated by medically orientated persons if placed in a familiar context. This is more so because specific problems arise in health sciences where the unit of interest is the living person and not some sort of abstract phenomena, objects, items or finance.

Statistics applied in biological sciences is simply called Biostatistics or Biometry meaning biological measurement. Many investigations in the biological sciences measured quantitatively with observations consisting of numerical information are called datum for singular and data for plural.

Statistics go under several names when applied to different fields. For example it is; Health Statistics when applied in Community Health or Public Health. Medical Statistics when applied in clinical medicine. Vital Statistics when applied in Demography that measures births, marriages and deaths. Biostatistics when applied to Biological Sciences.

Biostatistics has two main branches: Descriptive biostatistics concerned with summarizing and presenting data in an understandable form and inferential biostatistics that deals with drawing conclusions about a population based on observations made on a subset (sample).

3.2 Application of Biostatistics

The followings are some important applications and uses of Biostatistics as veritable tool in health and health-related areas:

1. In Disease Surveillance.

2. In Operation of Healthcare services such as; in the assessment of existing health conditions and their place in interventions considering the merits of various methods. In providing information on changing trends in health status in a given population.

3. In Public Health epidemiology, health services, healthcare policy and management.

4. In the evaluation of Healthcare services.

5. In decision-making e.g.in descriptive statistics, the researcher uses tables and diagrams to transform collected data to meaningful information. The researcher uses probability parameters to measure the level of certainty about an outcome. The researcher uses inferential statistics to draw conclusion about a large body of data by examining only a part (sample) of it.

6. In prioritization of Healthcare service implementation e.g. the application of Biostatistics helps in quantifying the magnitude of health and health-related problems and informs planning, implementation and further evaluation of such services and programs.

7. In comparison of health indices e.g. in defining what is normal or healthy in a given population and the limits of normal values; in finding the difference between means and proportions of normal values at two places or in different periods.

8. For the projection of health trends

9. For the analysis of health and health-related data

10. In Health Research e.g. in design and analysis of clinical and field trials in medicine e.g. testing vaccines, drugs or interventions and finding out if differences observed are statistically significant. In drug tests, e.g. finding the action of drugs, potency of new drugs and comparison of drugs. In epidemiological studies when the role of causative factors are tested. To find out association between attributes in disease causation. Testing the strength of associations and identified signs and symptoms of a disease or syndrome.

11. In population genetics, statistical genetics in order to link variation in genotype with a variation in phenotype. This has been used in agriculture to improve crops and animal husbandry e.g. animal breeding.

12. In ecological forecasting

13. In biological sequence analysis.

14. In systems biology for gene network inference or pathways analysis.

4.0 CONCLUSION

In this Unit we have explained the meaning and concept of Biostatistics. Adequate explanation was given in the variation in the nomenclature used in Biostatistics. Details of uses and appropriate application of Biostatistics in health and health-related fields have been provided in this unit.

5.0 SUMMARY

In this unit you have learnt that:

- biostatistics is all about ôthe Logic of Scientific reasoningô in uncovering the illusive truth
- biostatistics embraces those techniques pertaining to health and health-related sciences.
- specific problems arise in health sciences where the unit of interest is the living person and not some sort of abstract phenomena, objects, items or finance.
- The word statistics is derived from the Latin word ñstateò indicating the historical importance of governmental data gathering

- as biological entities are counted or measured, it becomes apparent that some objective methods are necessary to aid the investigator in presenting and analysing research data.
- biostatistics or Biometry is used in Biological Sciences.
- health Statistics is used in Community Health or Public Health.
- medical Statistics is used in clinical medicine.
- vital Statistics applies in Demography that measures births, marriages and deaths.
- there are various uses and application of Biostatistics in health and health-related fields.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define in detail what Biostatistics means.
2. Explain the concepts in Biostatistics.
3. What are the various nomenclatures used in Biostatistics and in which fields?
4. Explain the application of Biostatistics in public health.

7.0 REFERENCES/FURTHER READING

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Jos: Yakson Printing Press. Revised ed. 3-128.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Kano: Habason Nig. Limited. 32-34.

UNIT 2 DATA AND SOURCES OF DATA

CONTENTS

1.0 INTRODUCTION

In Descriptive Biostatistics, data is described in such a way that the classical features of the spread of the study population are highlighted for quick and easy visual understanding of its characteristics. It reduces the data to a smaller, concise and manageable size. It presents a summary of the spread of data. It does not draw inference or conclusion on the study population. The followings are used to organize data and describe the common characteristics of the population being studied, tables, diagrams, numerals. Each of these will reduce the data to a reasonable size either by compartmentalization or summary thereby aggregating the highlights or the striking features of the sample for better visual and easy comprehension of the spread of the data.

2.0 OBJECTIVES

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

- define and differentiate between raw data and population statistic
- state the common sources of data relevant to health and health-related fields.

3.0 MAIN CONTENT

3.1 Data

Data or population statistics are set of values of one or more variables recorded on one or more observational units. These are usually values resulting from counts and measurements. They are usually derived from biological sources such as medicine. Each of these is a datum. Raw data

are set of values that have not been statistically manipulated i.e. the information (data) generated is still in its natural form e.g. a set of 300 Level studentsô weight randomly collected during lecture will be considered as a set of raw data.

3.2 Sources of Data

3.2.1 Routine Sources

The routine sources of data could be from official publications or records collected by government agencies and institutions. It is natural that organizations keep records of day-to-day transactions of its activities. These are records that form secondary source of data in the sense that the investigator was not involved in the process of its collection or documentation. The records are there for him to access. It is important that in using such data its accuracy and reliability be considered. Examples of routine sources of data are:

- Vital statistics e.g. Birth, marriage and migration registrations.
- Published demographic data e.g. Census.
- Data from institutions such as; Hospitals viz. Out-patient and in-patient records, cancer registry etc. Dispensary and Health Post records. Schools enrolment or entry records etc. Industries viz. Pre-employment and periodic medical records, employment bio-data etc. Disease Surveillance and Notification system records.

3.2.2 Ad hoc Sources

This is a planned process of generating data either in the field or laboratory for a purpose. It constitutes a primary source because the investigator plans, supervises and controls the data collection. This may be in the form of:

- Field surveys.
- Laboratory research work.
- Experiments.

4.0 CONCLUSION

In this unit we have been able to discuss the differences between data and raw data and the common sources of data.

5.0 SUMMARY

In this Unit we were able to in details:

- define that raw data is different from statistic because its natural form has not been altered or undergone any statistical manipulation.
- identify the various common sources of data that are ad hoc and routine relevant to health and health-related fields

6.0 TUTOR-MARKED ASSIGNMENT

1. Define data and differentiate it from statistic.
2. List and classify the various common sources of data.

7.0 REFERENCES/FURTHER READING

Murray, R. S and Larry, J. S. (2006). Statistics. Theory and Problems of Statistics. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). The Basics in Biostatistics, Medical Informatics and Research Methodology. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Singha, P. (1996). An Introductory Text on Biostatistics. 2nd edition. Habason Nig. Limited. Kano: 32-34.

UNIT 3 METHODS OF DATA COLLECTION

CONTENTS

1.0 INTRODUCTION

It is pertinent to state here that the integrity of data collected will depend on the collectorôs expertise or skill and the instrumentôs validity.

2.0 OBJECTIVES

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

- state the various methods used in data collection
- explain how to use the various methods in data collection
- identify appropriate methods to be used for various data collection
- explain how to ensure the validity of methods used in data collection.

3.0 MAIN CONTENT

3.1 Use of Personal skills

Here basic instruments are not involved but personal expertise is brought to bear in:

- Observations
- Interviews e.g. In-depth interviews (IDIs), Focused Group Discussion (FGD), History taking.
- Physical examinations e.g. Inspection, Percussion, Palpation.
- Record reviews e.g. retrospective studies.

3.2 Use of Instruments

This is either with or without expertise:

- Use of instrument without expertise, e.g. using simple weighing scale to record weights. Much expertise or technicality is not required to do it. Here the instrumentôs validity will be an important factor of the integrity of the data collected.
- Use of instrument with expertise, e.g. using stethoscope to listen to heart beats for the presence of murmurs or microscope to examine blood film. Here expertise is required for accurate information and data generation. The researcherôs expertise and instrumentôs validity will be an important factor of the integrity of the data collected.

3.3 Use of Questionnaires

The questions are expected to be simple and non-ambiguous and free of double interpretations in meaning. The language used must be preferably that of the respondent. Number of questions for the respondent should not be cumbersome. Ensure uniformity of questions from one respondent to another.

3.3.1 Types of questionnaires

1. **Self-administered questionnaire:** The respondent collects the questionnaire and completes it independently before returning it.
    i. Advantages of self-administered questionnaire; Good for busy respondents, Convenient for researcher, does not require interviewers i.e. Saves Time and cost, Provides confidentiality.
    ii. Disadvantages of self-administered questionnaire; Completeness not assured, Loss of questionnaire, Information generated may not be reliable, Respondent may not understand some questions
2. **Interviewer administered questionnaire:** The respondent is guided by the field worker to respond to questions as the former fills in the information provided in the questionnaire.
    i. Advantages of interviewer administered questionnaire; authentic information generated. Completeness assured. Clarification provided.
    ii. Disadvantages of interviewer administered questionnaire; Lack confidentiality, Cumbersome, Requires interviewer e.g. Consumes more time and more expensive
    iii. Inconveniencing to respondent.

3.3.2 Types of questions asked

i. Open ended questions e.g. how much beer do you drink in a day? When do you go to bed? etc.
ii. Closed ended or pre-coded questions e.g. how many bottles of beer do you drink a day? 1-2 bottles, 3-4 bottles, more than 5 bottles, none at all, declined. At what time do you go to sleep? Before 8pm, 8-10pm, 10-12pm, after mid-night.

4.0 CONCLUSION

In this unit we have been able to discuss the common methods of data collection which are classified in personal skills, use of instruments and questionnaires.

5.0 SUMMARY

In this Unit we were able to:

- discuss the various personal skills as common method of data collection
- discuss the various common instruments used as methods for data collection
- mention types of questionnaires used in data collection
- describe kinds of questions asked in methods of data collection.

6.0 TUTOR-MARKED ASSIGNMENT

1. List and explain the various methods used in data collection.
2. Describe the various categories of questionnaires used in data collection.
3. Discuss the types of questions used in methods of data collection.

7.0 REFERENCES/FURTHER READING

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Habason Nig. Limited. Kano: 32-34.

UNIT 4 MEASURING INSTRUMENTS

CONTENTS

1.0 INTRODUCTION

Measuring instruments are varieties of instruments used in measuring variables especially in a living person and the commonest used instruments are the anthropometrical instrument. Appropriate measuring instruments with reliable validity should be used in collecting data otherwise the wrong value is obtained with serious consequences become inevitable.

2.0 OBJECTIVES

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

- explain the various types of measuring instruments in health and health-related fields.
- identify the appropriate measuring instrument for various data
- explain how to use the various measuring instruments to collect relevant data
- explain the advantages and disadvantages of each measuring instrument.

3.0 MAIN CONTENT

3.1 Measuring Instruments

Some of the commonly used measuring instruments especially in the anthropometric measurement of individuals are:

3.1.1 Weighing scales

These are used to measure the weight of the subject. Regardless of the type used, the important thing is to make sure it works properly and that it is calibrated correctly. Users should always follow the instructions that came with the weighing scale to ensure its accuracy. The subject to be weighed should be bare-footed with light clothing. There are various types of weighing scales and those of high precision should be used.

i. **Balance Scale:** Balance scales are weighing instruments that are upright with a platform where one stands on and utilizes a balancing weight on its top. The weight is moved on the top bar until the scale is balanced in the center to indicate the weight being measured. It requires space and it is cumbersome to carry for field work.

ii. **Scale:** Spring weighing instrument is the standard scale that is common and has been in use for long e.g. bathroom weighing scale. As one steps on the floor of the instrument the weight compresses the spring which then causes the disc to move. Gravity helps to determine how far the disc needs to turn and a spindle points to the weight being measured. Though the spring weighing scale is reliable excessive use and heavy weights weaken the spring and may give incorrect weight. It usually has a simple knob, button or other application for its recalibration after each use i.e. adjustment to zero point. The infant weighing scale is specially designed for safety as the infant is laid supine in it without clothes.

iii. **Digital Scale:** With advancement in technology came to be the digital weighing scales. One simply steps on the scale and the weight is calculated and projected on a visual screen. These weighing scales are usually powered and if by battery can read incorrectly when the battery is low. However, they have the added feature of a reset switch if there is a problem with the scale. Digital weighing scales offer additional features that can provide information on weight loss, Body Mass Index (BMI), muscle mass, water content and more. These additional features offer more detailed information all of which greatly affect ones overall weight. Weight can vary depending on the time of the day, body water levels, and for women, hormones, menstruation and even stress levels.

3.1.2 Height meters

These are instruments used to measure heights of individuals. Examples of such are:

i. **Tapes**: Tape height instrument is suitable in measuring individualôs height with an approximation of ±1 cm. It can be secured to the wall and measures up to 2 meters.Measuring tapes are also used in measuring the Mid Upper Arm Circumference (MUAC), Head and Abdomen circumferences and also in calculating, Waist- Hip ratio etc. Measuring tapes are space-saving instruments of measurement and allows for a quick reading.

ii. **Stadiometer**: These are stand-alone instruments used for adults because they can be upright in standing position. Commonly used Stadiometers are; H-101 Stadiometer. This is a mechanical, traditional height measuring device. Dual reading height rod measures in centimeters the height of the individual. The result can be read either on the level of the crown of the head or 40 cm lower to the approximation of 0.1 cm. Wall mounted and Digital wall mounted Stadiometers. These are Height measuring instruments with optoelectronic measurement system for children that can stand erect, young people and also adults.

iii. **Infantometer**: This is a height measuring instrument for infants. It uses a measurement board that is intended to measure the length of infants and small children (<2 years) because standing them erect may not be feasible.

3.1.3 Calliper

A Calliper is a properly calibrated measuring instrument useful in measuring the Skin fold thickness.

Note that there are some measuring instruments that have the combination of height meters and weighing scales in one piece. The disadvantage such combined instruments have is that they are relatively big and heavy for field work and also occupy space when in a room.

4.0 CONCLUSION

In this unit we have been able to discuss the common measuring instruments which are used in measuring data especially in a living person. These instruments mainly used to generate anthropometric data. It was explain that it is important that appropriate measuring instruments

with qualities of reliability and validity be used in collecting data otherwise the misleading values with serious consequences will be used to draw wrong conclusions.

5.0 SUMMARY

In this unit we have learnt:

- the various types of measuring instruments in health and health related fields.
- how to identify the appropriate measuring instrument for various data
- how to use the various measuring instruments to collect relevant data
- the advantages and disadvantages of each measuring instrument
- that personal expertise with or without any instrument is important in data generation.

6.0 TUTOR-MARKED ASSIGNMENT

1. List the various types of measuring instruments in health and health related fields.
2. Specify appropriate measuring instrument for various data
3. Describe how to use the various measuring instruments discussed
4. What are the advantages and disadvantages of each measuring instrument?
5. Why is personal expertise with or without use of instrument important in data generation?

7.0 REFERENCES/FURTHER READING

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers.*7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

MODULE 2 SCREENING TESTS AND VARIABLES

UNIT 1 DEFINING SCREENING TESTS

CONTENTS

1.0 INTRODUCTION

A screening test is used to separate from a large population of apparently healthy people who have a high probability of having the disease being investigated or studied. This provides a diagnostic work-up for those that have the disease for appropriate prompt treatment. Screening test is usually applied to groups of people who are apparently well with reference to one particular disease under investigation with a criterion and cut-off point used in classifying the subjects as positive or negative.

The extent to which the screening results agree with those derived by the more definitive tests provides a measure of **S*ensitivity*** and ***Specificity***. The characteristics of an instrument of data collection are ***Validity*** and ***Reliability***. The quality of a screening test instrument or procedure is measured by its validity. Data collection method or instrument is considered reliable if the same result is obtained from using the same method on repeated occasions. It is critical that data collection instruments or methods should have good validity and reliability attributes.

2.0 OBJECTIVES

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

- define a screening test

- define validity of a screening test
- define the reliability of a screening instrument
- define and calculate sensitivity of a screening instrument
- define and calculate the specificity of a screening instrument.

3.0 MAIN CONTENT

3.1 Validity

Validity is the ability an instrument has to indicate correctly the condition it supposes to measure. The key indicators of validity are Sensitivity and Specificity. These components are determined by comparing the results obtained by the screening test with those derived from some definitive diagnostic procedures.

- **Sensitivity.** This is the ability the measuring instrument has to identify correctly those who have the disease or condition. It is expressed as a percentage i.e.

$$\frac{\text{[illegible]}}{\text{[illegible]}} \times 100$$

- **Specificity.** This is the ability the measuring instrument has to identify correctly those who do not have the disease. It is also expressed as a percentage i.e.

$$\frac{\text{[illegible]}}{\text{[illegible]}} \times 100$$

The yield of a screening test is dependent on the sensitivity and specificity of the test and the disease prevalence. High risk populations are usually selected for screening, thus increasing the yield. Initial screening provides a prevalence estimate and subsequent screening an incidence value.

A screening test that is very sensitive will have fewer false negatives while a test that is very specific will have fewer false positive. This can be demonstrated in a tabular form as shown in table 1.

Table 1. Screening Test

SCREENINGTEST	CONFIRMATORY TEST		TOTAL
	Has disease	**No disease**	
Tested Positive	True Positive (a)	False Positive (b)	(a + b)
Tested Negative	False Negative (c)	True Negative (d)	(c + d)
TOTAL	**(a + c)**	**(b + d)**	**(a + b + c + d)**

The proportion of diseased persons the test classifies as positive i.e. has the disease.

Sensitivity = $\frac{a}{a+c}$

= [illegible]

The proportion of non-diseased persons the test classifies as negative i.e. not having the disease;

Specificity = $\frac{d}{b+d}$

= [illegible]

The proportion of positive tests that identify diseased persons i.e.
Predictive value of a positive test = $\frac{a}{a+b}$

= [illegible]

The proportion of negative tests which correctly identify non-diseased persons i.e.

Predictive value of a negative test = $\frac{d}{c+d}$

= [illegible]

The proportion of all tests which are correct classifications i.e.

Accuracy of the test = $\frac{a+d}{a+b+c+d}$

= [illegible]

False Positive Rate = $\frac{b}{b+d}$

False Negative Rate = $\frac{\square}{\square\square\square}$

Note the following relationships;

Specificity + the False Positive ratio = 1
i.e. $\frac{\square}{\square\square\square} + \frac{\square}{\square\square\square} = 1$

Therefore, if the specificity of a test is increased the false positive ratio is decreased.

Sensitivity + False negative ratio = 1
i.e. $\frac{\square}{\square\square\square} + \frac{\square}{\square\square\square} = 1$

Therefore, if the sensitivity of a test is increased the false negative ratio will be decreased.

- **Validity Cut-offs of Screening Tests**

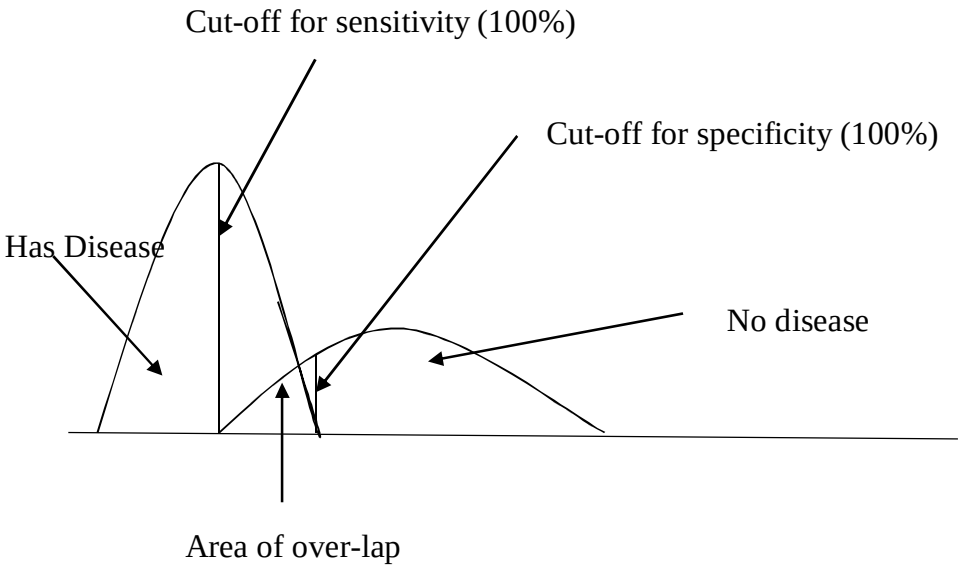


Fig. 1 Screening Test validity cut-off

This means that both the sensitivity and specificity are usually less than 100%. As efforts are made to increase sensitivity by reducing the area of over-lap more of **'No disease'** population area is included in the disease region and vice versa.

Note that the relationship between sensitivity and specificity is reciprocal to an extent.

- **The setting of cut-offs**

The setting of cut-offs for sensitivity and specificity will actually depend on several factors such as; Both the natural history of the disease and the effectiveness of early and late interventions must be known. If the disease has a latent period in development during which it is asymptomatic yet maybe detected by a screening test and prognosis, improved screening is then rewarding. If the disease is very rare, sensitivity must be high or else the few cases present will be missed. If the disease is very lethal and early detection markedly improves prognosis, high sensitivity is necessary. The cancers generally are such examples e.g. in cervical and breast cancer which are detected by cytology and mammography respectively. In these cases a proportion of false positive is tolerable, but false negatives are not otherwise it will be disastrous. However, in a prevalent disease, such as diabetes, for which treatment does not markedly alter outcome, specificity must be high and early cases may be missed but false positives are limited, otherwise the facility is overwhelmed by diagnostic demands on all the positives, both true and false.

3.2 Reliability

Reliability is the inherent variability of an instrument e.g. zero point fluctuation, unstable reagent. Reliability is also the replicability of the results e.g. a reliable instrument is one that gives consistent result.

Factors affecting the reliability of instrument are;

- Inherent variation of the instrument zero point fluctuation or unstable reagent.
- Fluctuation in the substance being measured e.g. a patient giving different answers to a question.
- Observer error; Inter-observer error. Error occurring as a result of different results from different observers on the same measurement; Intra-observer error. Error occurring as a result of an observer producing different results using the same method or instrument.
- Personal skills in the use of instruments. Here both human and instruments are involved in data collection e.g. reading of blood film for malaria parasites which involves both microscope and expertise.

4.0 CONCLUSION

In this Unit, we defined a screening test, validity of a screening test, the reliability of a screening instrument and how to calculate sensitivity and specificity of a screening instrument. We also discussed their cut-offs

and usage, advantages and disadvantages. We explained that validity and reliability of any screening test affects the credibility and consistency of the results. A good screening test is one with high sensitivity and specificity.

5.0 SUMMARY

In this unit we have learnt that:

- validity is the ability an instrument has to indicate correctly the condition it supposes to measure.
- the characteristics of a measuring instrument are validity and reliability
- the key indicators of validity are Sensitivity and Specificity.
- sensitivity and specificity are determined by comparing the results obtained by the screening test with those derived from some definitive diagnostic procedures.
- Sensitivity is the ability the measuring instrument has to identify correctly those who have the disease or condition expressed as a percentage
- specificity is the ability the measuring instrument has to identify correctly those who do not have the disease expressed as a percentage.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define validity of a measuring instrument.
2. What are the key indicators of validity?
3. Define and differentiate between sensitivity and specificity.
4. Define reliability.

7.0 REFERENCES/FURTHER READING

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Habason Nig. Limited. Kano: 32-34.

UNIT 2 VARIABLES AND CLASSIFICATION OF VARIABLES

CONTENTS

1.0 INTRODUCTION

Variables are attributes by which individuals differ from one another. They are quantities or measurements which vary in such a way that they may take any one of a specified set of values. Variables may be measurable in that integer or continuous values are assigned to them e.g. population or age. It is denoted as X and notation for orderly series as X_1, X_2, X_3,.. X_n, where n is a symbol for the last number in the series.

Variables could also be classified as non-measurable in that it is its attributes that can only be considered as its values e.g. sex, religion, treatment or surgery outcome.

2.0 OBJECTIVES

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

- define a variable
- identify various methods of classifying variables
- differentiate between various classes of variables
- give examples for each class of variables mentioned.

3.0 MAIN CONTENT

3.1 Variables

Variables are broadly classified into quantitative and qualitative variables.

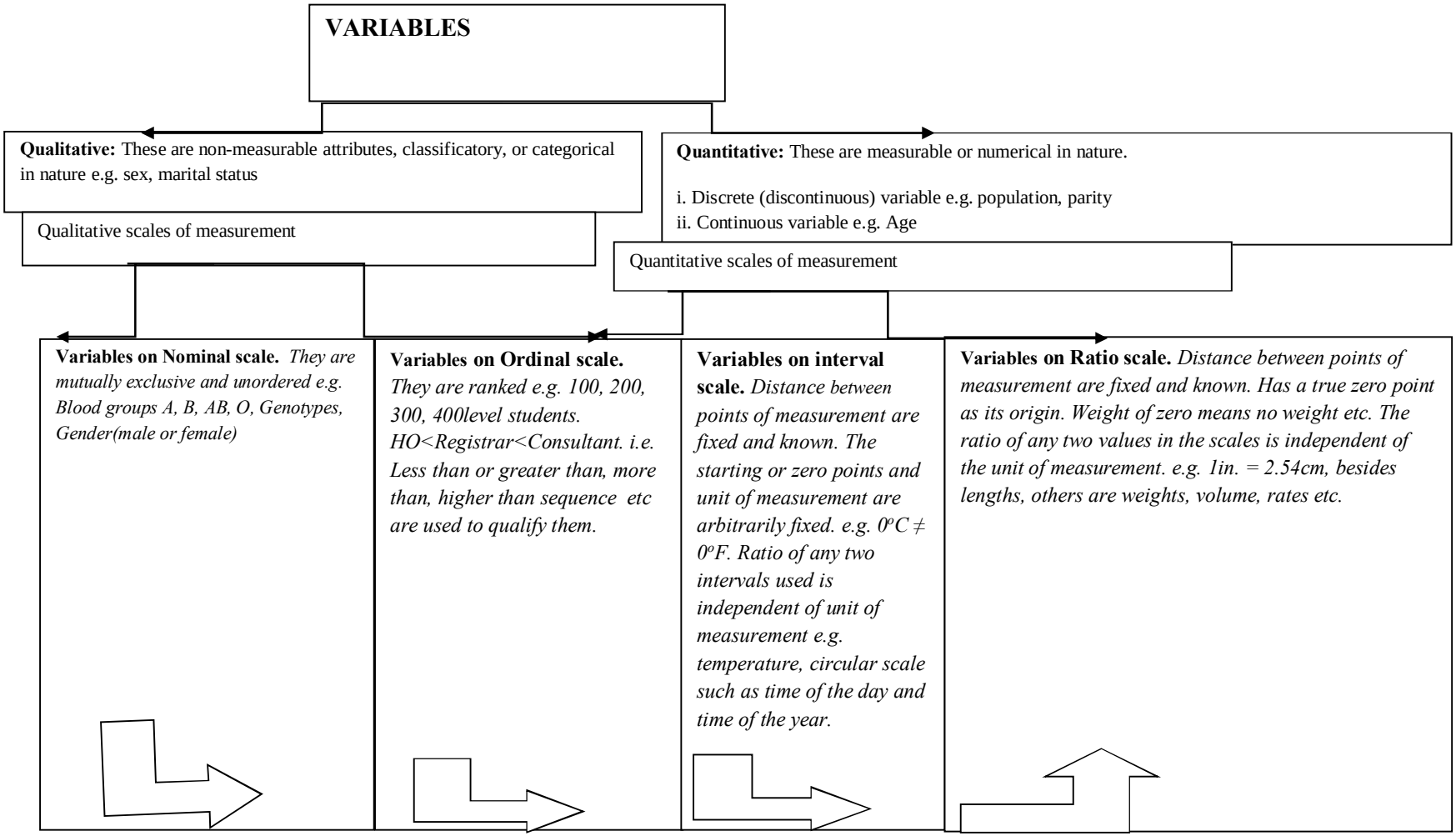
- **Quantitative Variables**

These are measurable or numerical variables in nature that are Discrete (discontinuous) variable such as population, parity or Continuous variable such as age, weight, height etc.

- **Qualitative Variables**

These are non-measurable variables that are attributes, classificatory, or categorical in nature such as sex, marital status, occupation, ethnic group etc.

Variables can further be classified as shown below;


VARIABLES
Qualitative: These are non-measurable attributes, classificatory, or categorical in nature e.g. sex, marital status
Quantitative: These are measurable or numerical in nature.
i. Discrete (discontinuous) variable e.g. population, parity
ii. Continuous variable e.g. Age
Qualitative scales of measurement
Quantitative scales of measurement
Variables on Nominal scale. They are mutually exclusive and unordered e.g. Blood groups A, B, AB, O, Genotypes, Gender(male or female)
Variables on Ordinal scale. They are ranked e.g. 100, 200, 300, 400level students. HO<Registrar<Consultant. i.e. Less than or greater than, more than, higher than sequence etc are used to qualify them.
Variables on interval scale. Distance between points of measurement are fixed and known. The starting or zero points and unit of measurement are arbitrarily fixed. e.g. 0°C ≠ 0°F. Ratio of any two intervals used is independent of unit of measurement e.g. temperature, circular scale such as time of the day and time of the year.
Variables on Ratio scale. Distance between points of measurement are fixed and known. Has a true zero point as its origin. Weight of zero means no weight etc. The ratio of any two values in the scales is independent of the unit of measurement. e.g. 1in. = 2.54cm, besides lengths, others are weights, volume, rates etc.

Features in the class of variables on the left are also found in the class of variables on the right as indicated by the arrows.

Fig. 2 Classification of variables

4.0 CONCLUSION

In this unit we have discussed variables as quantities or measurements that vary in such a way that they may take any one of a specified set of values. It was noted that variables may be measurable in that integer or continuous values are assigned to them such as in population or age. Again variables could be classified as non-measurable because its nature is attributory and not tangible and can only be considered in terms of its values without counts e.g. sex, colour etc.

5.0 SUMMARY

In this unit we have learnt:

- what variables are
- and identified the various methods of classifying variables
- how to differentiate between various classes of variables
- given examples in each class of variables mentioned
- discussed common features seen in some variables

6.0 TUTOR-MARKED ASSIGNMENT

1. Define a variable.
2. Identified the various methods in classifying variables.
3. Differentiate between various classes of variables.
4. Give examples variables in each class of variables discussed.
5. Mention common features identified in specified variables.

7.0 REFERENCES/FURTHER READING

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Habason Nig. Limited. Kano: 32-34.

MODULE 3 ORGANIZATION AND PRESENTATION OF DATA

UNIT 1 TABULAR PRESENTATIONS

CONTENTS

1.0 INTRODUCTION

Tabular presentation of data is the most frequently and easily used because irrespective of the nature of the data and variable, an appropriate tabular presentation can be constructed. The simplest form of table is a list while the more complex ones are the contingency or cross tabulated tables. Data presented could be as simple as individual observations or grouped in specified classes. Predetermined class intervals based on certain criteria are used in grouping the data. For simple tables only one variable is presented while when there are two or more variables then contingency tables are used. Every table should have a title describing what is being presented, especially, in terms of who and what.

2.0 OBJECTIVES

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

- list all the various tables used in data presentation
- describe each table discussed for data presentation
- mention the features of a good table
- construct all the tables discussed

- group data for tabular presentation
- know appropriate table to be used for specific data.

3.0 MAIN CONTENT

3.1 Lists

A list is the simplest form of a table consisting of two columns viz. The first column is observational unit giving identity number to the variable value. The second column is the value of the variable for that unit.

Table 2. A list of fifty 300 level students' weight

S/N	Kg.	S/N	Kg.	S/N	Kg.	S/N	Kg.	S/N	Kg.
1.	70	**11.**	47	**21.**	59	**31.**	53	**41.**	63
2.	54	**12.**	56	**22.**	80	**32**.	55	**42.**	76
3.	53	**13.**	73	**23.**	56	**33.**	63	**43.**	70
4.	76	**14.**	66	**24.**	54	**34.**	61	**44.**	66
5.	53	**15.**	73	**25.**	59	**35.**	49	**45.**	73
6.	56	**16.**	56	**26.**	61	**36.**	63	**46.**	63
7.	48	**17.**	56	**27.**	48	**37.**	57	**47.**	73
8.	55	**18.**	76	**28.**	61	**38.**	66	**48.**	76
9.	78	**19.**	47	**29.**	61	**39.**	63	**49.**	63
10.	85	**20.**	55	**30.**	53	**40.**	70	**50.**	78

Source: *Hypothetical data*

3.2 Frequency Table

This can be Ungrouped data table or Grouped data table. Ungrouped data frequency table is achieved by transforming Table 2 to frequency table as shown in Table 3. The table has three columns summarizing the data by showing the frequencies of the variables measured. The ungrouped data table is a technique for systematically arranging the data to indicate the frequency of each variable value. This reduces the size of the data especially when the raw data is large. It can be represented as a frequency table or histogram.
The columns are as follows:

- The first column is observational unit giving identity number to the variable value.
- The second column is the value of the variable for that unit.
- The third column is the corresponding frequency of each variable.

Table 3. Frequency distribution of the studentsô weight

S/N	Kg.	Tally marks	Freq.	S/N	Kg.	Tally marks	Freq.
1.	47	11	2	**10.**	61	1111	4
2.	48	11	2	**11.**	63	~~11111~~ 1	6
3.	49	1	1	**12.**	66	111	3
4.	53	1111	4	**13.**	70	111	3
5.	54	11	2	**14.**	73	1111	4
6.	55	111	3	**15.**	76	1111	4
7.	56	~~11111~~	5	**16.**	78	11	2
8.	57	1	1	**17.**	80	1	1
9.	59	11	2	**18.**	85	1	1

Grouped Data Frequency table is used when the variables are continuous or large the variable values may become infinite and need to be grouped. Using a frequency table without grouping the values will not make much difference from the original or raw data. Therefore a summary frequency table is produced by distributing the data into classes or categories and determining a fixed number of values that will be contained in each class. The following procedures are followed in grouping the values;

- Determine the range of values which is the difference between the largest and the smallest values.
- Decide on the number of classes by fixing class interval for the series of variable. Though arbitrarily fixed, most times it is usually determined by the size and form of data and certain requirements e.g. age classification following certain convention, stratification or milestone.
- Determine the width of the class intervals which has to be constant for all class intervals.
- Choose the upper and the lower limits of the class intervals such that each observation lies strictly within a class interval. This avoids ambiguities and ensures that no values lie on the boundary between two intervals.
- The class or group interval between the groups should not be too broad or narrow. The number of groups or classes should not be too many or few.

- List the intervals in order, considering each observation in turn and allocate it to the interval into which it falls. Use tally marks in obtaining the class frequencies as shown in table 3.
- The class interval should be the same throughout.

A grouped data frequency table also has three columns. This is achieved by transforming table 3 to a grouped data table as shown in Table 4.

- The first column contains the identity number of the class intervals.
- The second column is the class groups containing the variable values.
- The third column is the corresponding total frequency of values in each class interval.

Table 4. A grouped data frequency table

S/N	Class	Freq.
1.	40-49	5
2.	50-59	17
3.	60-69	13
4.	70-79	13
5.	80-89	2
Total		**50**

3.3 Relative Frequency Table

This is achieved by transforming group data table 4 to a grouped data relative frequency table as shown in Table 5.The relative frequency table has a fourth column that contains a proportion of the total frequency in each row either for the grouped or ungrouped frequency table. This is calculated by dividing the class frequency by the total frequency for all the classes and expressed as a percentage. It becomes a relative frequency distribution when the frequency column is replaced by the relative frequency column.

Table 5. A grouped data relative frequency table

S/N	Class	Freq.	Relative Freq. %
1.	40-49	5	10
2.	50-59	17	34
3.	60-69	13	26
4.	70-79	13	26
5.	80-89	2	4
Total		**50**	**100**

3.4 Cumulative Frequency Table

The grouped data relative frequency table is further transformed to a grouped data cumulative frequency table as shown in Table 6. The cumulative frequency table could either be for the grouped or ungrouped data. It has a column of frequencies that are cumulative from the first row in that the frequency at a value is the sum of the frequencies of the values less than or equal to that value.

Table 6. A grouped data cumulative frequency table

S/N	Class	Freq.	Cum. Freq.
1.	40-49	5	5
2.	50-59	17	22
3.	60-69	13	35
4.	70-79	13	48
5.	80-89	2	50
Total		**50**	

Table 7. A grouped data cumulative relative frequency table

S/N	Class	Freq.	Cum. Freq.	Cum. relative Freq. %
1.	40-49	5	5	10
2.	50-59	17	22	44
3.	60-69	13	35	70
a.	70-79	13	48	96
b.	80-89	2	50	100
Total		**50**		

3.5 Properties of a Good table

- Should have a title indicating at least who and what, i.e. person and the variable being measured or presented.
- Should be drawn to scale.
- Should have a key.
- Should be appropriately labelled.
- Should have uniform features.
- Should not be clumsy but simple and self-explanatory.

4.0 CONCLUSION

In this Unit we have been able to describe tabular presentation of data as an important method of data presentation that provide visual features of the population characteristics in compartmentalized form. Various types of tables were described and constructed with worked examples for

clearer understanding. The properties of a good table were also highlighted.

5.0 SUMMARY

In this Unit we have learnt that:

- tabular presentation of data is one of the methods used in descriptive Biostatistics in highlighting the characteristics of the spread of any population data
- there are several forms of tables ranging from a simple list to frequency and cross tabulated tables
- tables provide a visual understanding of proportions of the values of variables of population data being presented.

6.0 TUTOR-MARKED ASSIGNMENT

1. List the various tabular presentation of data.
2. Describe each table method mentioned.
3. Describe how the columns and rows of a table are determined and constructed.
4. What are the features of a good table for data presentation?
5. What is the usefulness of tables in descriptive Biostatistics?

7.0 REFERENCES/FURTHER READING

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in One Book. Revised Ed.Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Habason Nig. Limited. Kano: 32-34.

UNIT 2 DIAGRAMMATIC PRESENTATION OF DATA

CONTENTS

1.0 INTRODUCTION

The nature of the data collected and its variables determine the type of diagram that will be used to describe the population spread. The followings are the various types of variables and the types of diagrams appropriate for their presentation. For Qualitative or Categorical Variables the followings are some appropriate diagrams that should be used; Bar Chart, Pie Chart or Sector diagram, Pictogram or Picture diagram, Map diagram or Spot map. For quantitative or continuous variables the following diagrammatic representations can be used to present quantitative variables; Line Chart or Graph, Cumulative frequency diagram or Ogive, Frequency Polygon or percentage Ogive, Histogram, Dot graph or Scatter diagram

2.0 OBJECTIVES

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

- state the various types and categories of diagrams used in descriptive Biostatistics
- state the properties of a good diagram
- state the advantages and disadvantages of each diagram
- identify and know the appropriate diagrammatic presentation for various category of variables
- successfully construct the various type of diagrams.

3.0 MAIN CONTENT

3.1 Bar Chart

Bar Charts are graphic presentation of categorical variables drawn in rectangular forms with their lengths proportional to the frequencies or magnitudes of the variable represented. The Bar Charts can be in the form of two or more bars that are drawn adjacent to each other for comparison. The individual barôs frequencies should sum up to per cent. It is a comparison of bars by their frequencies of varying columns and shades.

- **Single Bar Chart**

A single Bar Chart indicating the frequencies of each group variables of the data in Table 4.

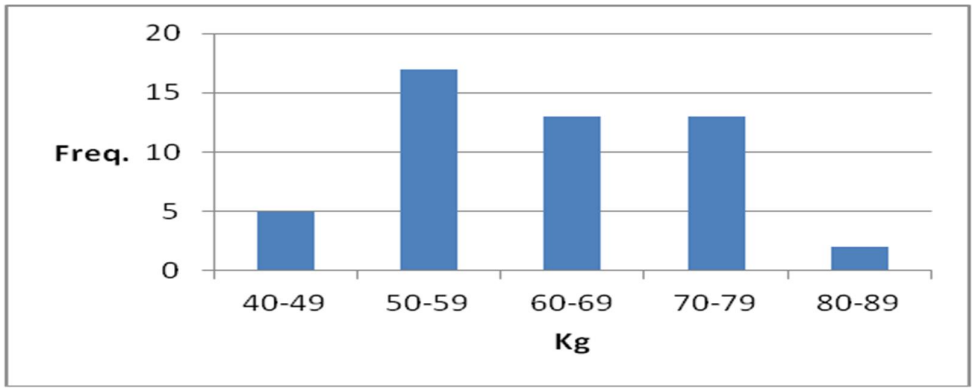


Figure 3a. A Single Bar chart representing each frequency.

- **A Segmented Bar Chart**

Single bar chart sub-divided into segments representing each group of variable shown in Table 4.

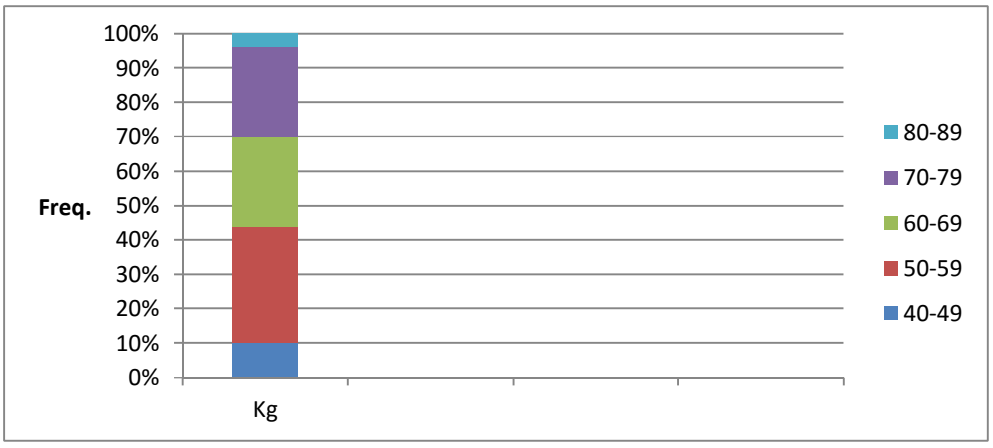


Figure 3b. A Component or Composite Bar chart each segment representing each frequency.

- **Contiguous Bar Chart**

Another method of diagrammatic data presentation of categorical variable on bar Charts is bars drawn on opposite directions. Single bars are drawn in opposite directions i.e. above and below the zero line and also in an increasing and decreasing order respectively with the frequency of each bar compared with the frequency of its opposite bar. Each bar can represent a categorical variable such disease entity or an outcome e.g. in an epidemic of a disease, opposite bars are drawn below and above the zero line. Each will then represent the number of deaths

(below bar with negative signs) while the survivors (above bar with positive signs) from affected individuals. Each contiguous bar will represent the successive years being studied e.g. 2008, 2009, 2010, 2011, 2012 and 2013 for the five bars as shown in fig. 5. This will give a visual interpretation of the trend of such epidemic over the five years as well as the survival rate.

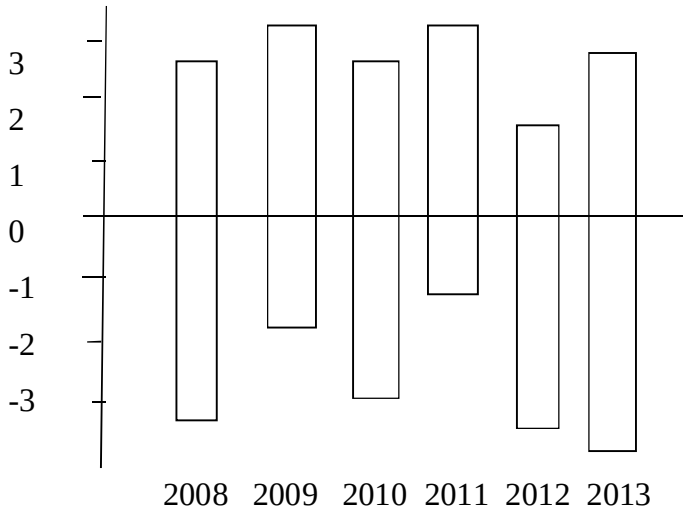


Fig. 3c. Ascending and descending bars

3.2 Histogram

This is a diagrammatic presentation of a continuous quantitative variable in rectangular form proportional to the class frequency. The area of each rectangle is proportional to the frequency of observations in each class. Ideally the width of each rectangle is equal and fixed representing the class interval of the variable.

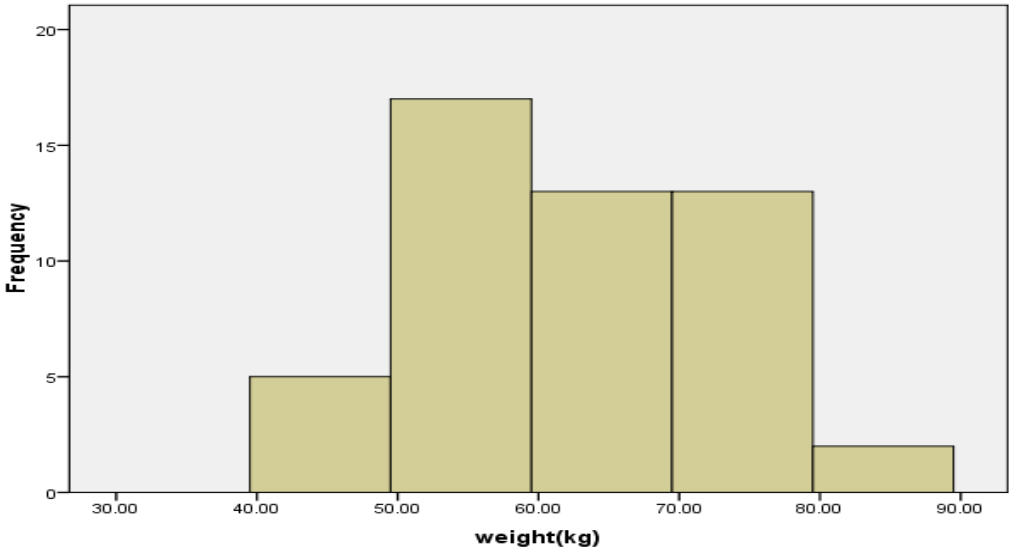


Fig. 9. Continuous quantitative variable frequency

3.3 Pie Chart

This consists of a circle whose area within the circle represents the total frequency. The total area is then divided proportionally into various segments to represent various variables. The data in Table 5 is here transformed into segments in a Pie Chart.

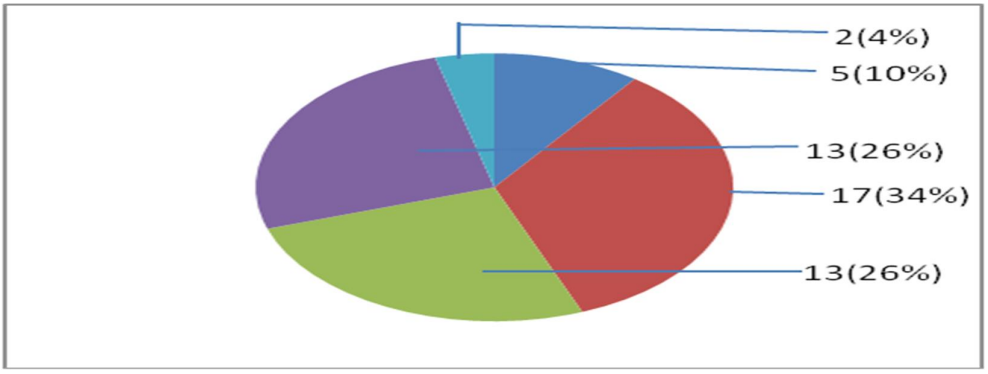
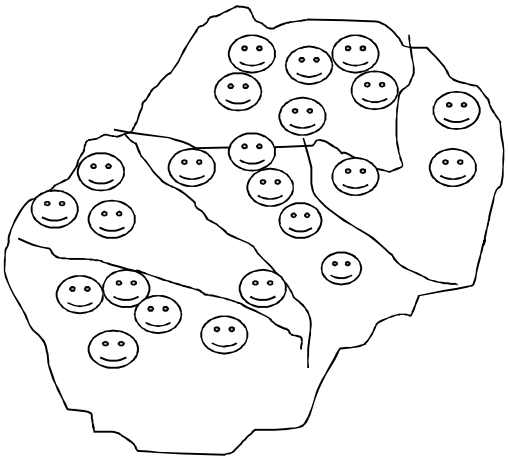


Fig. 4. Distribution of categorical variables

3.4 Pictogram

Also known as cartogram. This involves the use of drawings or symbols to represent diagrammatically the various variables of interest. A unit value of the variable should be represented by a standard symbol or drawing to depict its magnitude or frequency.



= 1,000 yearly live births in each state of the country.

Fig. 5. Using symbols to represent live births.

3.5 The Summary of the nature of data and choice of appropriate presentation

Table 8: Data Presentation Summary

Data and tabulation		
S/N	**Nature of Data**	**Tabular method of presentation**
1.	Small data set e.g. < 20	List or simple frequency table.
2.	Individual observations many in number involving only one variable.	Frequency tables. Grouped and ungrouped.
3.	Individual observations involving two or more variables.	Cross-tabulated tables.
Tables and Diagrammatic methods		
S/N	**Tabular Data**	**Diagrammatic methods**
1.	Frequency tables of quantitative variables of one set of data.	Histogram, Frequency polygon.
2.	Frequency tables of quantitative variable of two sets of data.	Frequency polygon.
3.	Frequency tables of categorical data or discrete variables.	Bar or Pie Chart
Nature of data and diagrammatic methods		
1.	Categorical or Classificatory data	Bar Chart Pie Chart

2.	Numerical data	Pictogram Dot graph (scatter diagram) Line graphs Histogram Frequency polygon Ogive

4.0 CONCLUSION

In this Unit we have been able to discuss and describe in details the various types and categories of diagrammatic presentation of data to visually appreciate the characteristics of the population spread. It reduces the data to a compartmentalized and concise form that the highlights are better appreciated visually. This can aid in formation of hypothetical statement on the features observed in the population described but not with a conclusion. We have also been able to explain how appropriate diagram for the right variable should be used.

5.0 SUMMARY

In this unit we have learnt that;

- diagrams are used in descriptive statistics to reduce the size of data by compartmentalization for highlighting of the outstanding features of the population
- there are various diagrams for appropriate variables which should be taken into consideration
- the nature of the data collected and its variables determine the type of diagram that will be used to describe the population spread.
- bar charts, Pie Chart or Sector diagram, Pictogram or Picture diagram, Map diagram or Spot map are used for qualitative of categorical variables.
- line Chart, Graph, Cumulative frequency diagram or Ogive, Frequency Polygon or percentage Ogive, Histogram, Dot graph or Scatter diagram are used for quantitative or continuous variables.

6.0 TUTOR-MARKED ASSIGNMENT

1. What are the uses of diagrams in descriptive Biostatistics?
2. What are the various diagrams used in descriptive Biostatistics?
3. Match diagrams against the appropriate variables for data collected.

4. List the appropriate diagrams used for qualitative or categorical variables.
5. List the appropriate diagrams used for quantitative or continuous variables.

7.0 REFERENCES/FURTHER READING

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers.*7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill.New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Habason Nig. Limited. Kano: 32-34.

UNIT 3 MAPS AND GRAPHS

CONTENTS

1.0 INTRODUCTION

Graphs could be polygonal or dot graphs in which successive points relative to the horizontal axis are linked by straight lines. This occurs when the variable on the X-axis is continuous and there is at most a single value on the Y-axis corresponding to any value on the X-axis. If the Y-axis represents the cumulative or relative frequency it is called a frequency polygon. If the X-axis represents the cumulative or relative frequency it is called a time series.Map diagram or Spot Mapare map drawings that are used to indicate delineated areas such as communities, LGAs, States and the frequency of the variable being measured indicated in each of such locations. This is similar to pictogram.

2.0 OBJECTIVES

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

- define and describe various types of Map diagrams and Graphs used if presenting data
- identify rightly appropriate Map diagrams and Graphs for the right data presentation
- discuss the features of good Map diagrams and Graphs for data presentation
- construct Map diagrams and plot various types of graphs
- state the differences between various graphs
- discuss the comparative advantages and disadvantages of the diagrams described.

3.0 MAIN CONTENT

3.1 Relative Frequency

A graph of Relative Frequency drawn from data in Table 4

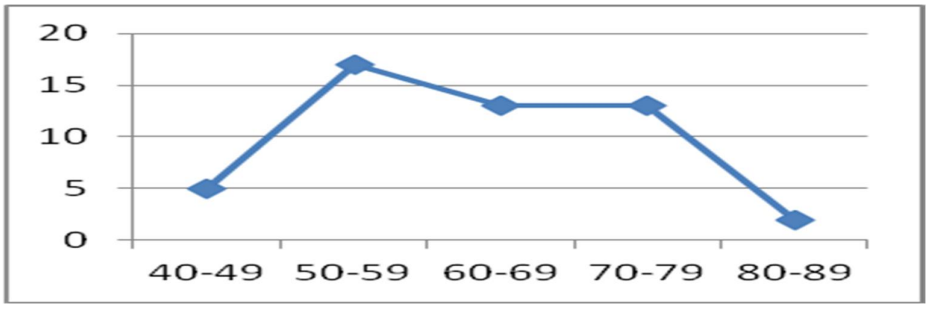


Fig. 6. Categorical variables showing class frequency intercepts

3.2 Ogive

A graph of cumulative frequency polygon is also called an Ogive drawn from the 4th column of the data in Table 6.

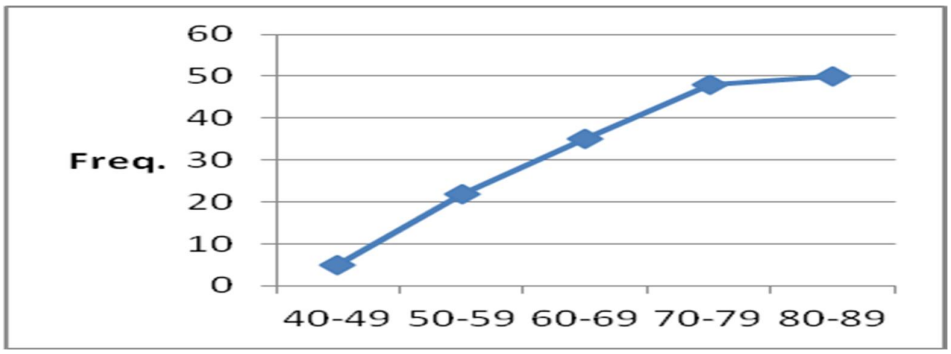


Fig. 7 Categorical variables showing cumulative class frequency intercepts

3.3 Percentage Ogive

A graph of cumulative relative frequency polygon is called a percentage Ogive drawn from the data in the 5th Column in Table 7.

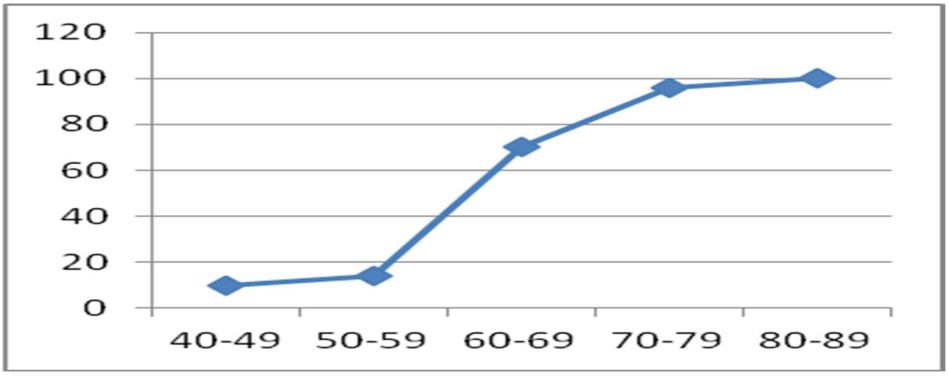


Fig. 8. Categorical variable showing cumulative relative frequency intercepts

3.4 Frequency Polygon

When the midpoint of the top of each rectangle of the histogram in Fig. 11 is connected a frequency polygon is produced. The total area under the curve will be equal to the sum of the rectangles.

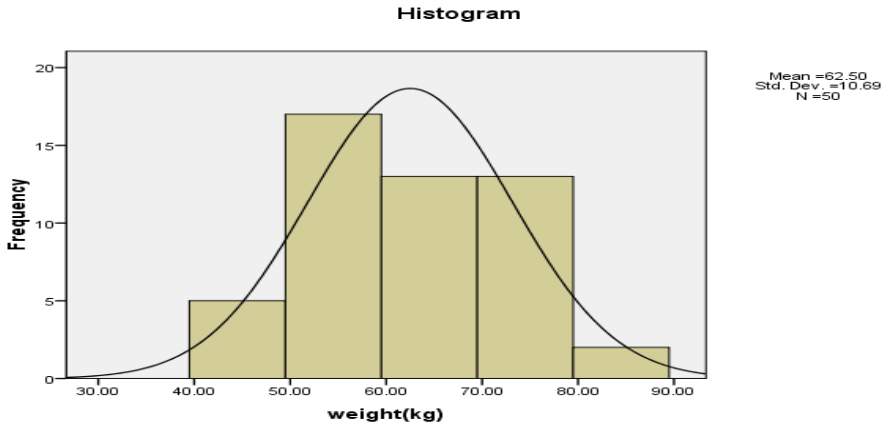


Fig. 10 Continuous quantitative variable showing midpoint frequency intercepts

3.5 Scatter diagram or Dot Graphs

Here each observation in Table 2 is represented by a point corresponding to its value on horizontal axis as abscissa and vertical axis as ordinate. Note that the intercepts i.e. dots here are only 18 and not 50 because those with same frequency are clustered on one intercept.

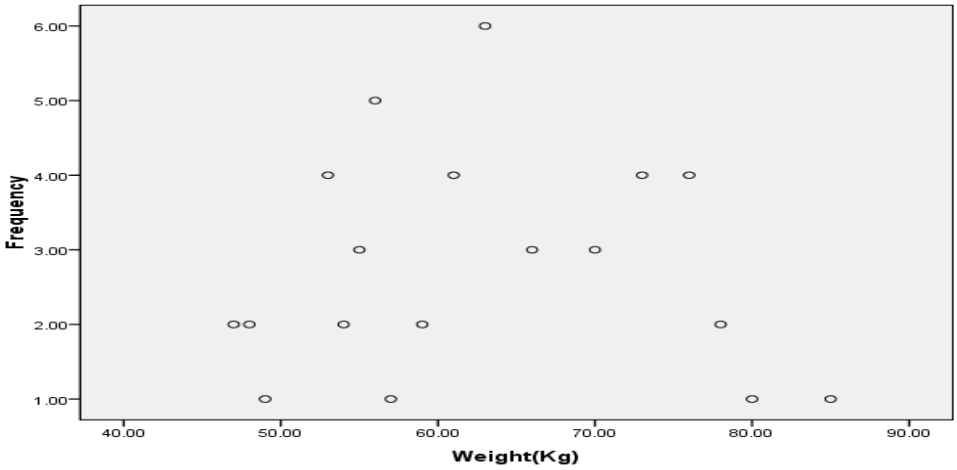


Fig. 11. Continuous quantitative variable frequency

- **One dimensional.** Here each observation is represented by a point corresponding to its value usually on horizontal axis only.

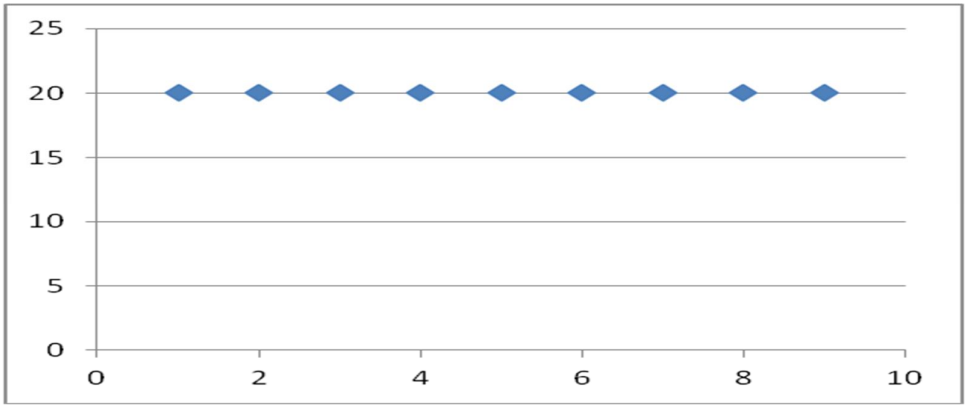


Fig. 12. One dimensional scattergram

4.0 CONCLUSION

In this unit we were able to define and describe various types of Map diagrams and Graphs used if presenting data. Attempts were made to identify rightly appropriate Map diagrams and Graphs for the right data presentation and the features of a good Map diagrams and Graphs for data presentation explained. You were also taught how to construct Map diagrams and plot various types of graphs and to know the differences between various graphs. It was explained to know the comparative advantages and disadvantages of the diagrams described in order to choose best table at any given time for data presentation.

5.0 SUMMARY

In this unit you have learnt:

- how to define and describe various types of Map diagrams and Graphs used in presenting data
- how to identify rightly the appropriate Map diagrams and Graphs for the right data presentation
- the features of good Map diagrams and Graphs for data presentation
- to construct Map diagrams and plot various types of graphs
- the differences between various maps and graphs
- the comparative advantages and disadvantages of the diagrams described.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define and describe the various types of Map diagrams and Graphs for data presentation.
2. Identify the appropriate Map diagrams and Graphs for defined data presentation.
3. What are the features of a good Map diagram and Graph used for data presentation?
4. Describe how you will plot Map diagrams and various types of graphs discussed.
5. What are the differences between the various types of graphs discussed?
6. What are the advantages and disadvantages of the various tables?

7.0 REFERENCES/FURTHER READING

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Edition. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition: Habason Nig. Limited. Kano. 32-34.

MODULE 4 NUMERICAL MEASURES

Unit 1 Measures of Central Tendency
Unit 2 Measures of Location
Unit 3 Measures of Dispersion or Variability

UNIT 1 MEASURES OF CENTRAL TENDENCY

CONTENTS

1.0 INTRODUCTION

Measures of central tendency are also known as averages. These generally indicate where the centre of the data spread lies. The most popular and frequently used are; Mean, Median and Mode. However each has its peculiar use and application and none could be of much use and application without relating it the measure of variability of the population spread such as standard deviation or coefficient of variation.

2.0 OBJECTIVES

At the end of this Unit, you should be able to;

- define and understand the use of mean as a measure of central tendency
- know and understand the various types of mean
- define and understand the use of median as a measure of central tendency
- define and understand the use of mode as measures of central tendency
- know how to calculate the mean, median and mode
- understand the advantages and limitations in the use mean, median and mode as measures of central tendency

3.0 MAIN CONTENT

3.1 The Mean

- **Arithmetic Mean**

This measure implies arithmetic average or mean which is commonly denoted as $\bar{x}$ i.e. pronounced as x-bar. This is the sum of all observations divided by the total number of observations in an ungrouped data and given as;

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

The Arithmetic mean in table 1 will be

$$\bar{x} = \frac{\text{[illegible]}}{\text{[illegible]}} = 62.52$$

In grouped data it is given as; $\bar{x} = \frac{\sum fx}{n}$

where $\sum$ is the summation of all the groups, f is the class frequency, x is the midpoint value in each group and n the total number of observations.

Characteristics and uses of Arithmetic Mean

- It is the centre of gravity of all values.
- In calculating arithmetic mean all observations are used.
- It is a stable average. Though affected by all values it is the least affected of all averages by fluctuations in samples.
- Sample means show less variation than individual values.
- It is not necessarily a value belonging to the group.
- Cannot be used when handling qualitative variables.
- Cannot be calculated if a single observation is missing.
- Cannot be calculated if extreme class is open i.e. no lower or upper limit e.g. < 20 or >50 class.
- Not a preferred average in skewed distribution.
- Variability of sample means depends on sample size.
- Its value is statistically useful when attached to standard deviation which may vary as shown in Fig. 13.
- It bisects a normal distribution into two equal halves.

Mean and standard deviation in a normal distribution

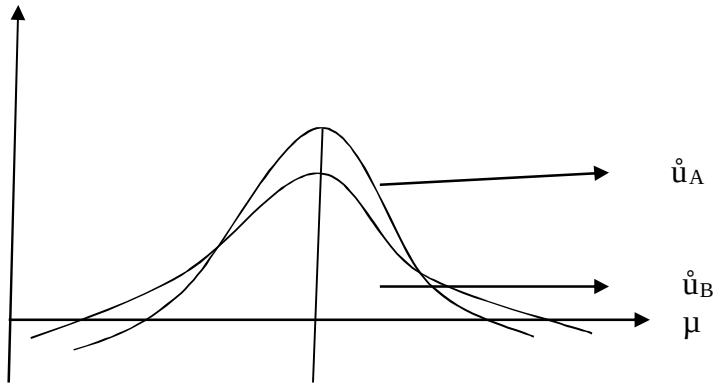


Fig. 13. Normal distribution curve

The two distributions have the same mean óµô but different standard deviations ůA and ůBi.e. the values measured in population B are more scattered and spread than in population A. So comparing their means without considering their standard deviations will be misleading.

- **Weighted mean ($\bar{x}_w$)**

In certain situations not all the observations will be measured precisely as others. Therefore a rational way out is to give relatively more weight to the more precise observations. Thus if the observations x_1, x_2, x_3.....,x_n have associated weights w_1, w_2, x_3,.....,w_n respectively then;

$$\bar{x}_w = \frac{\text{[illegible]}}{\text{[illegible]}}$$

- **Harmonic mean ($\bar{x}_H$)**

Thisis the average of all the arithmetic means of various populations or groups of measurement. If the various arithmetic means of various groups are given as $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$,.....$\bar{x}_n$, then;

$$\bar{x}_H = \frac{\text{[illegible]}}{\text{[illegible]}} \quad \text{or} \quad \bar{x}_H = \frac{\text{[illegible]}}{\text{[illegible]}}$$

- **Geometric Mean ($\bar{x}_g$)**

Thisis the nth root of the product of all observations

$$\bar{x}_g = \sqrt[n]{\text{[illegible]}}$$

Characteristics and uses of Geometric Mean

- Geometric mean is zero when any observation is zero.
- Geometric mean is imaginary when any observation is negative irrespective of the magnitude of other observations.
- Geometric mean is used in calculating population growth rate.

3.2 Median

The median is the middle most value in a set of ordered observations that ends with an odd number. The average of the two most middle values is taken when the set of the ordered values end with an even number i.e.

- When n is odd, Median = $x_{(n+1)/2}$
- When n is even, Median = $\frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}]$
- When the data is grouped, Median = $L + \frac{\frac{n}{2} - F}{f}c$ where L is the value of the lower boundary of the group where the median falls, n is the total observations, F is the total frequency below the group where the median falls, c is the class interval and f is the class frequency of the group where the median falls.

Characteristics and uses of the Median

- It is used in irregular and skewed distributions.
- It is not affected by outliers.
- Can be calculated if extreme class is open i.e. no lower or upper limit e.g. < 20 or >50 class.
- Can be used while dealing with qualitative variables.
- It is a better index for describing average number of cases of communicable diseases which have well defined cycle.
- It is also a better average in describing the trend of an illness over a period of time including epidemic years.
- Cannot be determined directly for even number observations.
- It is not based on all observations.

3.3 The Mode

Mode is the value that has the highest frequency i.e. most occurring value in a set of values. In a discrete probability distribution it is the value at which its probability mass function takes its maximum value or at the peak i.e. the value that is most likely to be sampled.

- **Types of Mode**

The distribution is unimodal if only one value has the highest frequency and bimodal if two values. For more than one value with highest frequencies, it is considered multimodal.

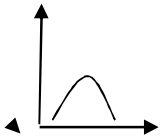


Fig. 14a. Unimodal distribution.

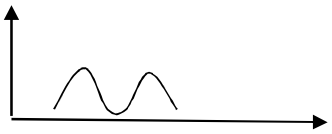


Fig. 14b. Bimodal distribution

Characteristics and uses of the Mode

- Does not exist if all values in a set occur with the same frequency.
- Not affected by outliers.
- It gives the point at which the observations cluster and converge.
- May have multiple values (mode).
- Not based on all observations.
- Used when the most repeated variable is wanted.
- Affected by fluctuation of sampling.

4.0 CONCLUSION

In this unit we discussed that measures of central tendency are also known as averages that generally indicate where the centre of the data spread lies. It was explained that the most popular and frequently used measures of central tendency are the Mean, Median and Mode. It was noted, however each has its peculiar use and application and none could be of much use and application without relating it the measure of variability of the population spread such as standard deviation or coefficient of variation.

5.0 SUMMARY

In this unit we learnt;

- how to define and understood the uses of mean as a measure of central tendency
- the various types of mean as used as measure of central tendency
- the use of median as a measure of central tendency
- the use of mode as measures of central tendency
- how to calculate the mean, median and mode

- the applications, advantages and limitations in the use of mean, median and mode as measures of central tendency

6.0 TUTOR-MARKED ASSIGNMENT

1. Define and explain the use of mean as a measure of central tendency.
2. Mention the various types of mean as used as measure of central tendency.
3. Define and explain the use of median as a measure of central tendency.
4. Define and explain the use of mode as measures of central tendency.
5. What are the various types of Mode?
6. Explain how to calculate the mean, median and mode.
7. What are the applications, advantages and limitations in the use of mean, median and mode as measures of central tendency?

7.0 REFERENCES/FURTHER READING

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book.Revised Ed. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Habason Nig. Limited. Kano: 32-34.

Syvia, Wassertheil-Smoller. (1990). *Biostatistics and Epidemiology. A primer for health professionals*. Springer-Verlag. New York: 119.

Wayne, W. D. (2006). *Biostatistics. A Foundation for Analysis in the Health Sciences*. 7th edition. John Wiley and Sons. New Delhi: 57-71.

UNIT 2 MEASURES OF LOCATION

CONTENTS

1.0 INTRODUCTION

Measures of location are measures that locate points on the entire range of the distribution of a variable. Commonly used measures of location are; Quartiles, Quintiles, Deciles, Percentiles.

- **Q_1= First Quartile** i.e. The frequency between first quartile and origin is 25% of total frequency.
- **Q_2= Second Quartile** i.e. The frequency between second quartile and origin is 50% of total frequency.
- **Q_3= Third Quartile** i.e. The frequency between third quartile and origin is 75% of total frequency.

2.0 OBJECTIVES

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

- define and understand the various measures of location
- explain the concept and usage of measures of location
- know how the respective measures of location were derived
- understand the advantages and limitations in the use of measures of location
- explain the comparative advantages and limitations in the use of the measures.

3.0 MAIN CONTENT

3.1 Quartile Deviation or Semi-inter Quartile range

There are three different points located on the entire range of the variable. The total frequency is divided into four equal parts with the lower quartile with25% observations below 1st quartile and 75% observations above 3rd quartile.

Quartile Deviation= $\frac{(\square\square\square\square)}{\square}$

3.2 Quintiles

This has four locations of 5equal segments with 20% observations below 1st quintile and 80% observations above 4th quintile.

3.3 Deciles

Divides the total frequency into 10 equal segments. There are nine locations with 10% of the observation below 1st decile and above 9th and between successive deciles.

3.4 Percentiles

Percentiles divide the total frequency into 100 equal segments. This has 99 locations with 1% of the observations below 1st percentile and 99% observation above 99th percentile.

4.0 CONCLUSION

In this unit, we have been able to define and explain measures of location as parameters that point to entire range of the distribution of the variables in the population spread. The concept and use of these measures have also been explained. Various categories of the measures of location have been described and used to demonstrate how the measures are derived. It was explained that their usage should be informed by taking advantage of their comparative advantages over the other measures as each of them has some disadvantage and limitations.

5.0 SUMMARY

In this unit we have learnt;

- the definition and understood the various measures of location
- the concept and usage of measures of location

- how the respective measures of location were derived
- the advantages and limitations in the use of measures of location
- the comparative advantages and limitations in the use of the measures.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define and explain the measures of location.
2. Explain the concept and usage of measures location.
3. List the measures of location.
4. Write down the various frequencies and locations of the measures of location?
5. What are the advantages in the use of measures of location?
6. What are the disadvantages and limitations in the use measures of location?

7.0 REFERENCES/FURTHER READING

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications, Ltd. Malden: 3-15.

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Edition. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

UNIT 3 MEASURES OF DISPERSION OR VARIABILITY

CONTENTS

1.0 INTRODUCTION

Measures of dispersion or variability show how spread out data are from the centre and are to each other in that, the population means may be the same values but different levels of variability from the mean. The measures of central tendency such as average value will not describe the degree of variation, spread or how scattered the observations of the distribution are without the accompaniment of the measure of dispersion. Commonest use of measures of dispersion is Standard deviation.

2.0 OBJECTIVES

At the end of this Unit we will be able to;

- define and conceptualize the measures of dispersion
- list the measures of dispersion
- explain the usefulness or otherwise of the measures of dispersion
- learn the various formulae used in measures of dispersion
- calculate and interpret values derived from the measures of dispersion
- explain the comparative advantage of the various measures.

3.0 MAIN CONTENT

3.1 Standard Deviation (SD or σ)

Commonest use of measures of dispersion is Standard deviation. Although Standard deviation is sensitive to outliers, it is the most

commonly used and useful measure of dispersion for the following reasons;

- It uses every observation in the distribution.
- It is the square root of variance.
- It is the best measure of dispersion.
- It summarizes the deviations of a large distribution in one value and unit of deviation.
- It is mathematically manageable with simple formula.
- Its unit of measurement/dimensionality is that of the original observations.
- It is used in calculating standard normal deviate (Z) which is measured in terms of Standard Deviations (SDs) and indicates how much an observation is bigger or smaller than the mean in units of SD.

$$Z = \frac{x - \mu}{\sigma}$$

For ungrouped data Standard Deviation is given as, $\sigma = \sqrt{\frac{\sum(x - \mu)^2}{n}}$

Where, x is the individual values, μ is the mean and *n* the sample size.

For grouped data Standard Deviation is given as, $\sigma = \sqrt{\frac{n\sum f x^2 - (\sum f x)^2}{n^2}}$

Where, f is the frequency of the individual groups, x is the midpoint value of individual groups and n the sample size.

3.2 Variance (SD^2 or σ^2)

Variance is the mean squared deviations from the mean value presented as the square of Standard Deviation. It is the arithmetic mean of the squared deviations.

Variance for ungrouped data is, $\sigma^2 = \frac{\sum(x - \mu)^2}{n}$

Variance for grouped data; $\sigma^2 = \frac{n\sum f x^2 - (\sum f x)^2}{n^2}$

3.3 Coefficient of variation

It is useful in comparing variations between two population variables that are dissimilar but may have the same characteristics. It is also used in situation when any change in the measurement brings about change in

the standard deviation in the same way as it does for the mean. It is useful in testing for homogeneity of independent populations for a certain feature. An example is in trying to know if variation in blood pressure is more in the older population than in the younger population. Coefficient of variation is expressed in percentage therefore independent of units of measurement. It reduces measure of dispersion to a dimensionless quantity.

CV = $\frac{\square\square}{\square}$ x 100

Example: In two series of women and men their mean systolic blood pressures were 128±10mmHg and 126±8mmHg respectively. Find which of the two populations shows greater variation.
Solution.

CV of BP for women = $\frac{\square\square}{\square\square\square}$ x100 = 7.8%

CV of BP for men = $\frac{\square}{\square\square\square}$x100 = 6.4%

Therefore, systolic blood pressure shows greater variation among the women than in men as $CV_{women} > CV_{men}$

3.4 Range

A range defines the normal limits of biological characteristics. In medical practice the normal range covers observations falling within 95% confidence limits. Range is simply obtained by subtracting the smallest from the largest value. i.e.

Range = $X_{(n)}$ ï $X_{(1)}$.

Although easy to obtain it does not use majority of the observations and therefore dependent on number of observations. It is easily affected by outliers i.e. extreme values because it uses only two extreme values.

- **Inter-quartile range:** This measures the difference between two quartiles distribution of 3rd quartile and 1st quartile. Half of this range is the **semi-interquartile range or quartile deviation.**

3.5 Mean Deviation

This is derived by adding up all the differences between the individual observations and then ignoring their signs and dividing it by the total observations. Mean Deviation application in statistics is not of much use.

Mean Deviation = $\frac{\square^{\circ}\square\square\ \square^{\circ}}{\square}$

3.6 Standard error of mean

Standard error of mean is also known as the standard deviation of a sampling distribution. It tells how the sample means are spread around the population mean. Therefore, it is the Standard Deviation of the sampling distribution of the mean of an infinite number of samples, each of size ñNò i.e. various means of samples drawn randomly from the population.

Note that Standard Deviation is a measure of the spread of the observations and should be reported with the estimated mean value when the aim is to describe the data distribution. While Standard error of the mean relates to the precision of the estimated mean and should be reported when attention is centred on the mean e.g. when one mean is compared with another.

Uses of Standard error of the mean

- To determine the probability of obtaining a mean which is a specified number away from the population parameter of interest.
- To determine the probability that the population parameter of interest falls within a given range of point around the sample statistics.

3.7 The Difference between Standard Deviation and Standard Error

For mean values to be meaningful they are often presented as $\square\pm$ SD and read as mean + or ï 1standard deviation. Sometimes means are also presented as $\square\pm$ s.e. and read as mean + or ï 1standard error. Therefore, standard deviation and standard error being used arbitrarily but they serve quite different functions. In order to understand the concept of standard error, one must consider that the purpose of statistical applications is basically to draw inferences from sampled populations or data by estimating their true mean.

Suppose we draw several units or samples from the same population the various means calculated from each sampling will vary slightly (hopefully) from each other if the same method is applied. Therefore, these sample means will form a normal distribution. Some of them will be very close to the true population mean while others will be on either sides of it. Therefore, this distribution will have its Standard Deviation which is from sample means rather than individual values. The standard

deviation of this distribution of sample means is called the standard error of the mean.

Thus the larger the sample size the better is our estimation of the true population mean. The standard deviation describes how disperse or variable the values are while the standard error is used to draw inferences about the population mean from which we have a sample.

3.8 Applications of measures of Dispersion

1. For Individual Observations;
   - Standard Deviation
   - Variance
   - Coefficient of variation
   - Range
   - Inter-quartile Range
   - Mean Deviation
2. For Samples;
   - Standard error of mean
   - Standard error of difference between two means
   - Standard error of proportion
   - Standard error of difference between two proportions
   - Standard error of correlation coefficient
   - Standard deviation of regression coefficient

4.0 CONCLUSION

In this unit we learnt that the measures of dispersion or variability show how spread out data are from the centre and are to each other in that, the population Means may be the same values but different levels of variability from the Mean. The measures of central tendency such as average value will not describe the degree of variation spread or how the observations of the distribution scattered are without the accompaniment of the measure of dispersion. The commonest used of measures of dispersion is Standard deviation because it uses all the items for the computation and retains the original unit of measurement or dimensionality.

5.0 SUMMARY

In this unit you have learnt:

- the measures of dispersion or variability
- how to define and conceptualize the measures of dispersion
- the uses of the measures of dispersion

- the various formulae used in measures of dispersion
- how to calculate and interpret values derived from the measures of dispersion
- the comparative advantage of the various measures of dispersion
- for the individual observations we use; the Standard Deviation, Variance, Coefficient of variation, Range, Inter-quartile Range, Mean Deviation
- for Samples we use; Standard error of mean, Standard error of difference between two means, Standard error of proportion, Standard error of difference between two proportions, Standard error of correlation coefficient, Standard deviation of regression coefficient.

6.0 TUTOR-MARKED ASSIGNMENT

1. List the measures of dispersion or variability.
2. Define the measures of dispersion.
3. What are the uses of the measures of dispersion?
4. State the various formulae used in measures of dispersion.
5. Explain how to calculate and interpret values derived from the measures of dispersion.
6. What are the comparative advantages of the various measures of dispersion?
7. What are the drawbacks of the measures of dispersion?

7.0 REFERENCES/FURTHER READING

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

MODULE 5 MEASURES OF RELATIONSHIP AND PROBABILITY

UNIT 1 MEASURES OF RELATIONSHIP

CONTENTS

1.0 INTRODUCTION

So far we have considered quantitative data in terms of averages, locations, variability and frequency distributions or curves. However, there are quantitative variables that are continuous in nature either in series or same persons that can be compared to establish not only any association but also the degree of the relationship. Therefore, measures of relationship are used in establishing the relationship between variables in terms of one influencing the other as in dependent and independent variables.

Scientific studies often require a description of the relationship between two variables. Usually in such circumstances we think of one variable as being influenced by the other. It has become conventional to denote the dependent variable i.e. the one being influenced by óYô and the independent variable by óXô. We are interested in describing the association between X and Y. To do this we have to measure jointly both X and Y on a series of subjects.

The simplest way of describing the relationship between X and Y is by plotting a scatter diagram. To construct it, the levels of the dependent variable Y is plotted against the corresponding levels of the independent

variable X for each subject. The resulting scatter diagram showing points of intercept indicate how Y varies with differing levels of X.
Although the scatter diagram is very useful for gaining visual impression of the relationship a more quantitative description is often needed. The major measures of relationship are;

2.0 OBJECTIVES

At the end of this Unit we will be able to;

- define and conceptualize the measures of relationship
- list the measures of relationship
- mention the usefulness or otherwise of the measures of relationship
- learn the various formulae used in measures of relationship
- calculate and interpret values derived from the measures of relationship
- construct a scatter gram to describe the relationship between dependent and independent variables
- state the comparative advantage of the various measures of relationship.

3.0 MAIN CONTENT

3.1 Linear Correlation

In order to determine the relationship between two or more variables measures of relationships are applied to the statistical units. The strength of this relationship can be measured by the use of óCoefficient of correlationô. The value may attain any numerical value from -1 to +1.
Negative Correlation means negative value and that the variables move in opposite directions i.e. when one increases the other decreases and vice versa. Positive Correlation means positive value and that the variables move in the same direction i.e. when one variable increases the other increases and vice versa.

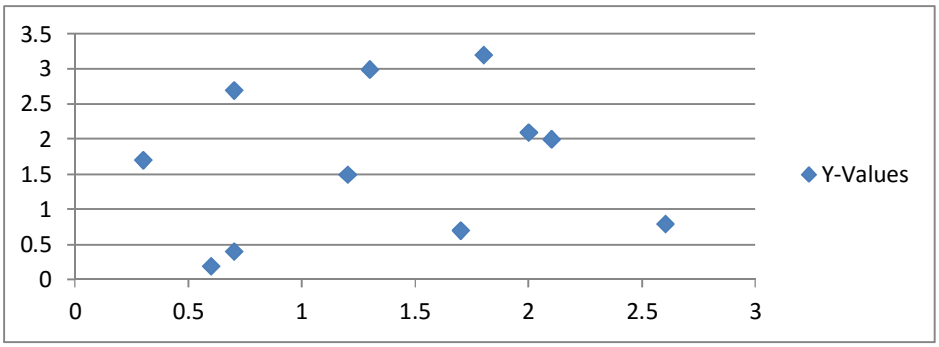
Scatter diagram is the simplest way to represent a bi-variate distribution e.g. measuring the heights and weights of group of people. The relationship of the two variables x and y are interdependent (co-related) and the measure of the degree of the closeness is estimated mathematically by the sample statistic órô thus;

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Therefore the properties of ór ô are;

- The magnitude of ór ô indicates the strength of the linear relationship between x and y.
- Two independent variables are uncorrelated. i.e. r; $_{xy=}$ 0.
- But two uncorrelated variables may or may not be independent i.e. r; $_{xy=}$ 0 implying that there is no linear relationship.
- Correlation coefficient lies within -1 and +1; i.e. -≤ r ≤ +1
- Correlation coefficient is independent of change of origin and scale.

An example of uncorrelated relationship is shown in Fig. 15 below e.g. height variations of adults in a normally distributed population and educational status.



r = 0

Fig. 15. A scatter diagram of no correlation

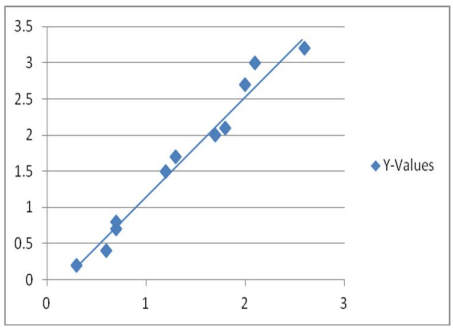


Fig. 16. (r = +1)

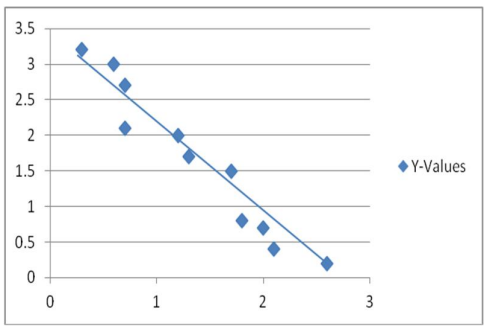


Fig. 17. (r = -1)

Scatter grams showing positive (fig. 16) and negative (fig. 17) linear correlations. An example of positively correlated relationship is age and weight distribution of under five children in a given population. Therefore, with increasing age (X-axis) there is corresponding increase in weight (Y-axis) as typically seen in under fives growth Chart. An example of a negatively correlated relationship is visual acuity and age

in a given population. As in adulthood with increasing age there is corresponding diminishing of individualôs visual acuity.

3.2 Spearman's rank correlation coefficient

Spearmanôs rank correlation coefficient is applied when one or more of the variables do not come from a bivariate normal distribution or at least one of the variables is measured on an ordinal scale i.e. ranked or requires a measure which does not depend on approximation to linear regression. Most commonly applicable in less complex distribution is the Pearsonôs correlation coefficient.

3.3 Pearson's correlation coefficient

When the relationship between variables is normally distributed such as in measures of height and weight the correlation coefficient is known as Pearsonôs correlation coefficient. There are three basic formulae that can be used in calculating Pearsonôs correlation coefficient ór ô viz.

- Using the variables and their mean values (x, y, $\bar{x}, \bar{y}$) $r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$

- Using only the variable values (x, y) $r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{(\sum x^2 - \frac{(\sum x)^2}{n})(\sum y^2 - \frac{(\sum y)^2}{n})}}$ $\frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$

- **Worked example**

Determine if there is a relationship between the weights and heights of school children generated from a normally distributed population shown below.

Weight (Kg)	12	18	22	40	48	50	52	53	55	62
Height (m)	0.94	1.52	1.2	1.5	1.56	1.54	1.58	1.6	1.61	1.74

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

X y	11.28	27.36	26.40	60.00	74.88	77.00	82.16	84.80	88.55	107.88
x^2	0.88	2.31	1.44	2.25	2.43	2.37	2.50	2.56	2.59	3.03
y^2	144	324	484	1600	2304	2500	2704	2809	3025	3844

Ɛxy = 640.31, Ɛx^2 = 22.36, Ɛy^2 = 19,738, $r = \frac{\text{□□□.□□}}{\text{□□□.□□}}$ = 0.96, DF = 10-2 = 8,

Critical value at 0.05 = 2.31

To test for significant relationship use, $t = \frac{\sqrt{(n-2)}}{\sqrt{(1-r^2)}} = \frac{2.236\sqrt{(10-2)}}{\sqrt{(1-0.90^2)}} = 5.5$

The critical value at 0.05 and DF = 8 is 2.31. Therefore, the relationship is statistically significant i.e. 2.31<5.5 or P<0.05.

- **Interpreting the magnitude of Correlation**

The following can serve as a general guide to interpret the magnitude of the correlation coefficient;

R	**Degree of correlation**
0.00	No Correlation
0.00 - 0.25	Weak correlation
>0.25 ï 0.5	Moderate correlation
>0.5 ï 0.75	Strong correlation
>0.75 ï 1.00	very Strong correlation

3.4 Linear Regression

In correlation it was the consideration of the direction and degree of the relationship between two continuous quantitative variables whereas in Regression it is the prediction of the values of one variable from known values of other variables. The relationship between any two variables can be plotted as a scatter diagram. The dependent variable is on the vertical/ordinate axis (Y) while the independent variable on the horizontal/abscissa axis (X). A mathematical model can be derived to represent the relationship thus;

$y = a + bx$

óyô is the dependent variable, óxô is the independent variable, óaô and ó bô are constants estimated by the method of least square and also ó bô is the slope of the regression of y on x.

The letters a and b (the regression coefficient) are calculated as follows from the data:

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \quad \text{or}$$

if you have 'r' and standard deviations of y and x, then $b = r\frac{s_y}{s_x}$

$a = \bar{y} - b\bar{x}$

□is the mean of all y-values and□ is the mean of all x-values.

Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of original units of the data. The line of regression is obtained by the principles of least square. For a simple linear regression used to predict a quantitative outcome variable (dependent) based on another quantitative explanatory variable (independent), the equation y=□ + □□is used based on the assumptions that;

- The form of regression equation is truly linear in relationship.
- The distribution of y values for each x is normal.
- The pairs of observations are all independent of each other.
- The variances of y distribution are the same for each value of x.

3.5 Coefficient of Association (Q)

The coefficient of association between variables can be illustrated by considering the association between measles outbreak and measles vaccination among infants as shown in table 9 a 2x2 2cross-tabulated table. Cell (a) are those who received measles vaccine and got infected, (b) are those who received measles vaccine but did not get infected, (c) are those who did not receive measles vaccine and got infected and (d) are those who did not receive measles vaccine and did not get infected.

- **Worked Examples**

Table 9.Distribution of measles out-break (Source: *Hypothetical)*

	HAD MEASLES		
	Yes	**No**	**Total**
Had measles vaccine	4(a)	192(b)	196
Never had measles vaccine	34(c)	113(d)	147
Total	**38**	**305**	**343**

Q = □□□□□□(□□□)
□□□□□□(□□□)

Q = □□□□□□□□(□□□□□□□)
□□□□□□□□(□□□□□□□)

Q = - 0.870

Note that the value of Q will vary from 0 ï 1. While 0 indicates no association, 1 indicates absolute or very strong association. While negative value indicates an association in opposite direction, positive value indicates an association in the same direction. Q can also be considered in ranges e.g.

0.00 ï 0.25 = Weak association
>0.25 ï 0.50 = Moderate association
>0.50 ï 0.75 = Strong association
>0.75 ï 1.00 = Very strong association

For Q = -0.870 in the above calculation, this value indicates that there is a very strong negative association between measles vaccination and infection. In order words with an increase in measles vaccination there is a decrease in measles infection in the studied population.

4.0 CONCLUSION

In this unit we have learnt there are quantitative variables that are continuous in nature either in series or same persons that can be compared to establish not only any association but also the degree of the relationship. Therefore, measures of relationship are used in establishing the relationship between variables in terms of one influencing the other as in dependent and independent variables.

5.0 SUMMARY

In this unit we have learnt;

- the definition and conceptualized the measures of relationship
- the usefulness or otherwise of the measures of relationship
- the various formulae used in measures of relationship
- how to calculate and interpret values derived from the measures relationship
- how to construct a scattergram to describe the relationship between dependent and independent variables
- the comparative advantage of the various measures of relationship.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define measures of relationship.
2. What are the various formulae used in measures of relationship?
3. What is the place of scatter gram in measures of relationship?
4. Differentiate between dependent and independent variables.
5. What are the comparative advantages of the various measures of relationship?

7.0 REFERENCES/FURTHER READING

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York: 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Dorling Kindersley Pvt. Ltd. Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics.* 2nd edition.Blackwell Scientific Publications, Ltd. Malden: 3-15.

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Murray, R. S and Larry, J. S. (2006). Statistics. Theory and Problems of Statistics. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

UNIT 2 PROBABILITY

CONTENTS

1.0 INTRODUCTION

For every conclusion there is a measure of uncertainty. This uncertainty is formally and numerically expressed as a probability. It measures the relative frequency of a particular event happening by chance in the long run. It is a branch of mathematics that helps us estimate the likelihood of something happening or not happening. Probability is the branch of mathematics that calculates the possible outcomes of given events together with the outcomesô relative likelihoods and distribution. Probability therefore is the relative frequency of occurrence with which an event is expected to occur or not to occur. The theory of Probability is the foundation of **Inferential Statistics.**

2.0 OBJECTIVES

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

- define probability and explain the concept and its application
- state the probability theories, laws and rules
- mention and state the formulae for the various Probabilities
- give examples of phenomena or events that follow a particular probability
- calculate and interpret probabilities values.

3.0 MAIN CONTENT

3.1 Definition and Concept of Probability

Probability theory is the foundation of statistical inference. This is so because inferential statistics deals with drawing conclusion about the

statistic of population from a sample drawn from it. It is important to always have in mind that this conclusion is always beclouded by an element of uncertainty as a result of the non-totality of the data collected. P[E] expresses the probability that the event óEô occurs and read as ñthe probability of Eò. Suppose an event óEô can occur in órô ways out of a total of ónô possible equal likely ways, therefore, the probability of occurrence of the event i.e.

Success óP[E]ô will be represented as;

$$P[E] = \frac{r}{n}$$

While failure will be represented as;

$$Pó[E] = 1 - \frac{r}{n}$$

3.2 A priori probability

A priori probability is the predicting the outcome of an event that is measureable. It is the proportion of times the event in a series of trials will occur e.g. suppose there are ónô equally likely outcome to a trial and an event óEô consists of n_Eó of these outcomes, then the probability of occurrence of óEô called its ósuccessô and denoted by óEô is,

$$P = Pr(E) = \frac{n_E}{n} = \frac{\text{number of favourable outcomes}}{\text{total number of possible outcomes}} \quad \text{i.e.}$$

Therefore the Probability of non-occurrence of óEô i.e. its failure is denoted by

$$q = \frac{n - n_E}{n} = 1 - p$$

A priori probability is one of such probabilities that form the basis of much of genetics e.g. in the determination of the sex and genotype of a baby as shown below.

- **Sex determination probability**

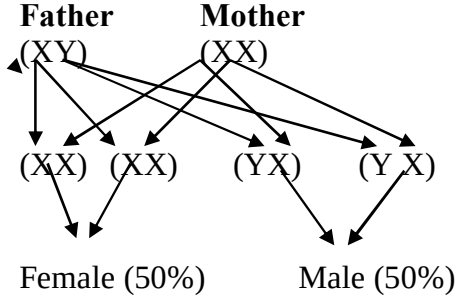


Fig. 18a. The probability of the childôs sex determination

- **Genotype determination Probability**

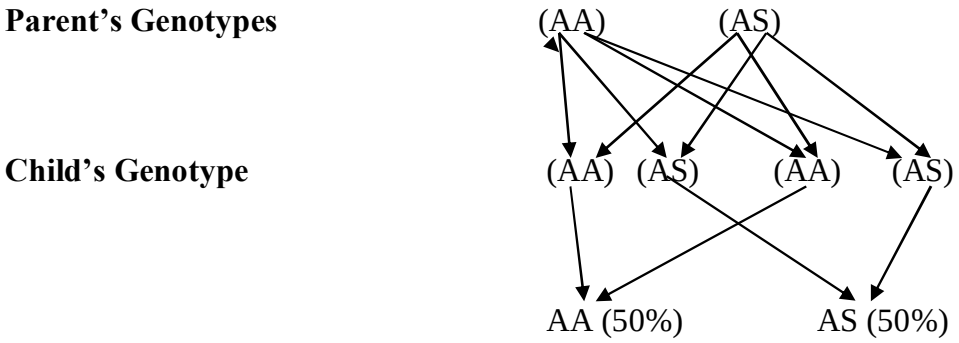


Fig. 18b. The probability of the childôs genotype determination.

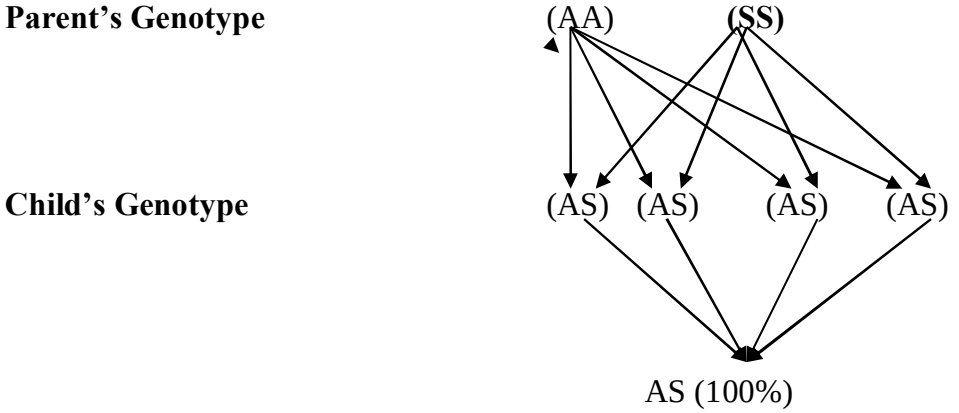


Fig. 18c. The probability of the childôs genotype determination.

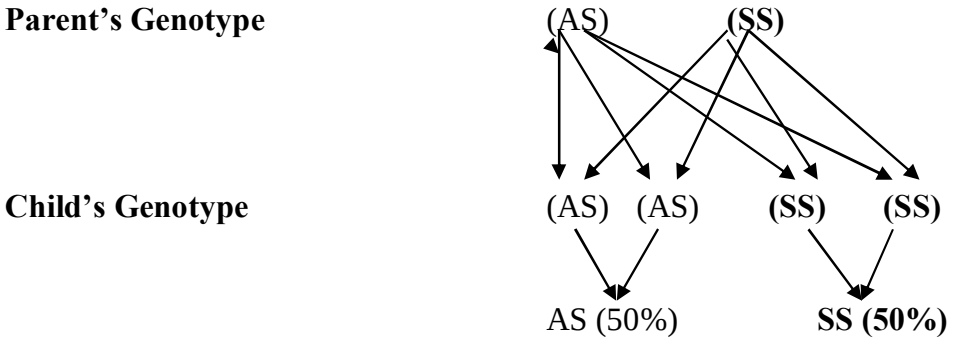


Fig. 18d. The probability of the childôs genotype determination.

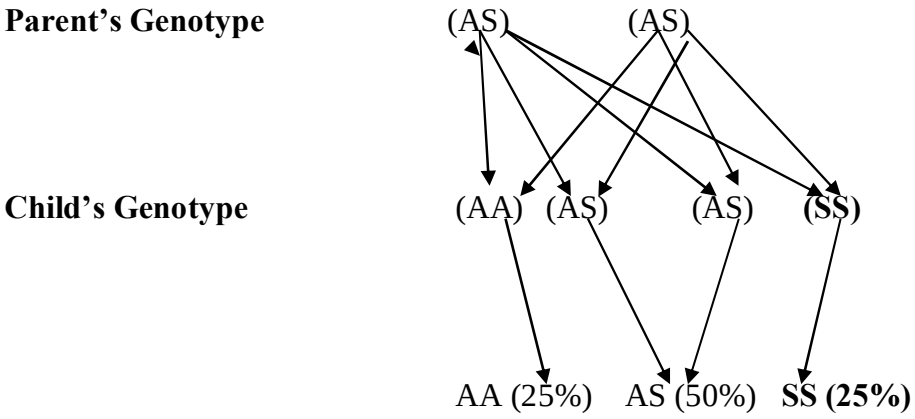


Fig. 18e. The probability of the childôs genotype determination.

3.3 Empirical probability

Empirical probability is a probability that considers the probability of an event pertaining to a procedure which depends on the relative frequency of occurrence of that procedure in a long series. It is an estimate that the event will happen based on how often the event occurs after collecting large data or doing an experiment in a large number of trials. It therefore approximates to A priori in an infinite repetition. It is based specifically on direct observations or experiences and calculated as;

$$P[E] = \frac{\text{□□ □□□□□□□□ □□□□□□□'□' □□□□□□}}{\text{□□□□□□□□ □□□□□□□□□□□□□ □□□□□□□□□□□}}$$

3.4 Subjective probability

Subjective probability is personalistic and based on intuition and cumulative experience. It is therefore, not measurable but explains the degree of oneôs belief and judgement. It varies from person to person and non-scientific.

3.5 Probability Thermometer

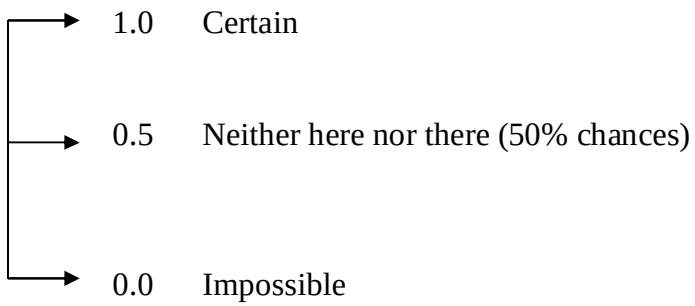


Fig. 19 Probability thermometer

If an event is certain to occur, the probability is a unit i.e. 1, while if the event is certain not to occur the probability is absolute zero. For any event óAô the probability of óAô occurring P(A) must be greater than or equal to zero and less than or equal to unit i.e.

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

The sum of the probabilities to all mutually outcomes of an event is a unit i.e.

$$\sum_{i=1}^{n} Pr(E_i) = 1$$

3.6 Probability Definition of Terms

- **Random Series:** This is the sequence of outcomes (events) in a very large number of trials.
- **Exhaustive Event/Cases:** This is the total number of possible events in any trial.
- **Independent events:** When the occurrence or non-occurrence of one event does not affect the probability of the outcome of the other. The probability of both occurring is the product of the probability of each given as $Pr(E_1E_2) = Pr(E_1)Pr(E_2)$
- **Dependent events:** When the occurrence or non-occurrence of one event affects the probability of the outcome of the other. When not necessarily independent we have $Pr(E_1E_2) = Pr(E_1)Pr(E_2/E_1) = Pr(E_2)Pr(E_1/E_2)$
- **Mutually Exclusive event:** Two events are mutually exclusive when the occurrence of one excludes the occurrence of the other.
- **Conditional Probability:** When the probability of the outcome of an event, óE_1ó is affected by the outcome of another eventóE_2ó then the probability of óE_1ô is conditional on óE_2ô. The conditional probability of óE_1ô given that óE_2ôhas occurred is $Pr(E_1/E_2)$ e.g. If after a social function 130 persons ate chicken out of which 85 had abdominal discomfort among the 350 that attended the function. The Probability of abdominal discomfort for those who ate the chicken expressed as conditional probability will be;

$$\mathbf{Pr} = \frac{85}{130} = 0.65 \text{ or } 65\%$$

3.7 Probability Rules

There are two basic rules guiding the calculation of all probabilities. These are;

1. **Addition Rule.** The addition rule is used for calculating the probability that at least one of the events E_1 or E_2 or both will occur e.g.

- Mutually Exclusive Events: For mutually exclusive events óE_1ô and óE_2ô the probability of either óE_1ó or óE_2ó occurring is the sum of their probabilities i.e. $Pr(E_1+E_2) = Pr(E_1)+Pr(E_2)$
- Events not necessarily Mutually Exclusive: When events óE_1ô and óE_2ó are not necessarily mutually exclusive the probability of either óE_1ô or óE_2ô or both óE_1ô and óE_2ô occurring is;
  $Pr(E_1+E_2) = Pr(E_1)+Pr(E_2)-Pr(E_1E_2)$

2. **Multiplication Rule**
   - For Independent events: If óE_1ô and óE_2ô are independent of each other, the probability of both occurring will then be the product of their probabilities i.e. $Pr(E_1E_2) = Pr(E_1)Pr(E_2)$.
   - Events not necessarily independent i.e.
     $Pr(E_1E_2) = Pr(E_2)Pr(E_1/E_2) = Pr(E_1)Pr(E_2/E_1)$

3.8 Mathematical or Classical Probability

Mathematical or Classical probability measures the number of ways that the event can occur divided by the total number of outcomes.

If in a trial outcome there are ón ô exhaustive, mutually exclusive and equally likely outcomes out of which ómô are favourable to occur in an event óEô with the probability;

$$P = P(E) = \frac{m}{n};$$

the probability of non-occurrence of the event E will be;

$$q = \frac{(n-m)}{n} = 1 - \frac{m}{n} = 1 - p$$

Therefore $p + q = 1$ and $0 \leq p \leq 1$.

Worked Example
If coronary heart disease case fatality is 15% and two patients are admitted into the Cardiac Unit, what is the probability that both or either will die?

Either will die (Addition rule) = 0.15 +0.15 = 0.30
Both will die (multiplication rule) = 0.15 x 0.15 = 0.0225

4.0 CONCLUSION

In this unit we learnt that probability theory is the foundation of statistical inference. This is so because inferential statistics deals with drawing conclusion about the statistic of population from a sample drawn from it. It is important to always have in mind that this conclusion is always beclouded with an element of uncertainty as a result of the non-totality of the data collected. This uncertainty is formally and numerically expressed as a probability. It measures the relative frequency of a particular event happening by chance in the long run. It is a branch of mathematics that helps us estimate the likelihood of something happening or not happening. Probability is the branch of mathematics that calculates the possible outcomes of given events together with the outcomesô relative likelihoods and distribution.

5.0 SUMMARY

In this unit we have learnt that:

- probability theory is the foundation of statistical inference
- conclusions are always beclouded by an element of uncertainty as a result of the non-totality of the data collected
- the uncertainty is formally and numerically expressed as a probability
- probability measures the relative frequency of a particular event happening by chance in the long run
- probability is a branch of mathematics that helps us estimate the likelihood of something happening or not happening
- probability is the branch of mathematics that calculates the possible outcomes of given events together with the outcomesô relative likelihoods and distribution
- probability is the relative frequency of occurrence with which an event is expected to occur or not to occur.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define Probability theory and explain how it is the foundation of statistical inference.
2. Why is it said that conclusions are always beclouded by an element of uncertainty?
3. How do we formally and numerically expressed uncertainty in statistical conclusion.
4. Define and explain Binomial theory.
5. Define and explain Poisson theory.
6. What is Mathematical or Classical Probability?
7. State Probability Rules.

7.0 REFERENCES/FURTHER READING

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York: 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Dorling Kindersley Pvt. Ltd. Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications Ltd. Malden: 3-15.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

MODULE 6 POPULATION DISTRIBUTION

UNIT 1 NORMAL DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

A normal distribution also known as Gaussian curve is one that is unimodal with the total area under the curve 100% or a unit. It is said to be symmetrical about μ with Mean, Mode and Median equal and lie on the same axis. It is characterised by population mean, μ and variance, $ů^2$ and for a constant ů a change in $μ_1$ to $μ_2$ shifts the curve along the x-axis to the right, if $μ_2 > μ_1$ and to the left when $μ_1 > μ_2$. For a constant μ a change in ů from $ů_1$ to $ů_2$ alters the peakedness of the curve. It is more peaked (taller or thinner) if $ů_1 > ů_2$ and less peaked (fatter or flatter) if $ů_2 > ů_1$. X-axis is an asymptote to curve while the intervals on either sides of μ encloses approximately a total probability of; 68.27% for 1 SD, 95.45% for 1.96 SD and 99.73% for 2.58 SD.

2.0 OBJECTIVES

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

- define and explain the concept in Normal Distribution
- mention and understand the properties of Normal Distribution
- discuss the importance of Normal Distribution or Gaussian curve
- discuss the uses and applications of Normal Distribution population.

3.0 MAIN CONTENT

3.1 Properties of a Normal Distribution Curve

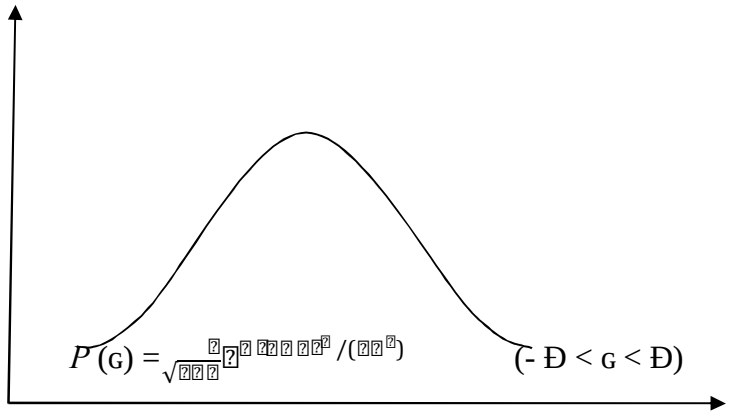


Fig. 21. Normal distribution (Gaussian curve)

- It is bell shaped i.e. unimodal
- Total area under the curve is 1 (100%) or a unit.
- Symmetrical about μ i.e. can be bisected into two equal symmetric halves.
- Mean, Mode and Median coincide (equal) i.e. lie on the same axis.
- Characterised by population mean, μ and variance, $ů^2$
- For a constant ů a change in $μ_1$ to $μ_2$ shifts the curve along the x-axis to the right, if $μ_2 > μ_1$ and to the left when $μ_1 > μ_2$.
- For a constant μ a change in ů from $ů_1$ to $ů_2$ alters the peakedness of the curve. It is more peaked (taller or thinner) if $ů_1 > ů_2$ and less peaked (fatter or flatter) if $ů_2 > ů_1$
- X-axis is an asymptote to curve.
- The intervals on either sides of μ encloses approximately a total probability of;

    - 68.27% for 1SD
    - 95.45% for 1.96SD
    - 99.73% for 2.58SD

Areas covered by each standard deviation about the mean are shown below.

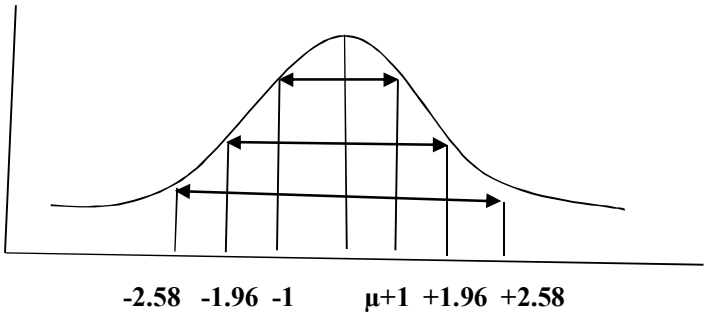


Fig. 22. Probability intervals under the normal distribution curve

3.2 Importance of Gaussian curve

- If variation is produced by large number of effects the distribution is normal.
- It fits most practical distribution occurring in real life and medicine i.e. Binomial and Poisson (P) can be approximated by normal distribution, provided that P is neither too close to zero nor unit and n is large.
- When the distribution is not normal, transformation techniques exist to make it normal.
- Sampling distributions of means and proportions are known to have normal distribution.
- Many distributions of sample statistic tend to normal for large samples and as such they can be studied with the help of normal distribution.
- If variables are not normally distributed, transformation techniques to make them normal exist.
- The entire theory of small sample tests such as, t, F and X^2 tests is based on the fundamental assumption that the parent population from which the sample is drawn follows a normal distribution.
- The statistical theory is elegant if assumptions hold.
- It is the cornerstone of all parametric tests of statistical significance.

4.0 CONCLUSION

In this unit we learnt that a Normal Distribution is also known as Gaussian curve. It is one that is unimodal with the total area under the curve 100% or a unit. The other features are symmetry with the Mean, Mode and Median equal and lie on the same axis. The population distribution is characterized by mean and variance and for a constant variance. Changing the Mean value shifts the curve along the x-axis.

When the Mean is constant a change in Standard deviation will alter the peakedness of the curve. That X-axis is an asymptote to the curve while the intervals on either sides of the Mean will enclose approximately a total probability of; 68.27% for 1 SD, 95.45% for 1.96 SD and 99.73% for 2.58 SD respectively.

5.0 SUMMARY

In this unit you have learnt:

- the definition of Normal Distribution and can explain the concept
- the properties of a Normal Distribution
- the importance of Normal Distribution
- the uses and applications of Normal Distribution population
- proportion of values covered on both sides of the mean for a given standard deviation.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define Normal Distribution and can explain the concept.
2. List and explain the properties of a Normal Distribution.
3. What is the importance of Normal Distribution?
4. What are uses and applications of Normal Distribution population?
5. Mention the percentage values covered on both sides of the mean for a given standard deviation.

7.0 REFERENCES/FURTHER READING

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Habason Nig. Limited. Kano: 32-34.

Syvia, Wassertheil-Smoller. (1990). *Biostatistics and Epidemiology. A primer for health professionals*. Springer-Verlag. New York: 119.

UNIT 2 STANDARD NORMAL DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

The standard normal distribution is one whose mean is = 0, Standard deviation is = 1 and the total area under the curve is = 1. Along the abscissa instead of x we have a transformation of x called the standard score Z. Thus the Z-score really tells us how many standard deviation from the mean a particular Z-score is. Any distribution of a normal variable can be transformed to a distribution of Z by taking each x value subtracting from it the mean of x and dividing this deviation of x from its mean, by the standard deviation.

One good thing about the Z distribution is that the probability of a value being anywhere between two points is equal to the area under the curve between those two points. Another important use of Z derives from the fact that we can also convert a sample mean (rather than just a single individual value) to a Z-score. Furthermore, the probabilities (areas) under the standard normal distribution are provided in statistical tables.

2.0 OBJECTIVES

At the end of this Unit, you should be able to;

- define and explain the concept in Standard Normal Distribution
- mention and understand the properties of Standard Normal Distribution
- discuss the importance of Standard Normal Distribution
- explain the difference and uses of Standard Normal Distribution and Z-Score

3.0 MAIN CONTENT

3.1 Standard Normal Distribution

Area of any segment under a standard normal curve has been calculated and available in a table. Transform data to standard normal curve before using the table. The transformation to standard normal curve is as follows;

$$Z = \frac{(\square\square\ \square)}{\square}$$

3.2 Standard Normal Distribution and Z-scores

In the Standard Normal curve the mean is taken as Zero and SD as Unit or One. Standard normal cure and Z-scores are used in comparing the values of a variable with those of reference curves. The analysis is carried out using the Z-scores rather than those of the original values. It is most commonly used in anthropometric data where growth charts are used to assess where an individualôs anthropometric value lies as compared to the standard.

Ideally, individualôs anthropometric measurements cannot be objectively interpreted unless they are related to the individualôs sex and age and compared to the same sex and age distribution of an appropriate reference population. An example of anthropometric reference population is the NCHS/WHO growth reference data. This helps to improve on the interpretation of collected raw data. A Z-score expresses how far a value is from the population mean and expresses this difference in terms of the number of standard deviations by which it differs. Therefore Z-score value is from the standard normal distribution i.e.

$$\text{Z-score} = \frac{(\square\square\ \square)}{\square}$$

$\square$ is an observed value, µ is the mean reference value and ů the standard deviation of the corresponding reference data.

Worked Example: The mean height of 500 students is given as 160±5cm.

- What are the chances of heights above 175cm being normal if their height follows normal distribution?
- What percentage of students will have height above 168cm?
- How many of the students will have height between 168cm and 175cm?

Solution 1.

$Z = \frac{\square\square\square\square\square\square\square}{\square} = \frac{\square\square}{\square} = 3$

This value 3, corresponds to 0.0013 on the Z-score table. Therefore 0.13% of the students will have height above 175cm which is the chance of being taller than 175cm.

Solution 2.

$Z = \frac{\square\square\square\square\square\square\square}{\square} = \frac{\square}{\square} = 1.6$

This value 1.6 corresponds to 0.0548 on the Z-score table. Therefore, 5.48% of the students have height above 168cm.

Solution 3.

The number of students having height above 168cm and below 175cm will be 0.0548 ï 0.0013 out of 1 i.e. Unit. Therefore, only 5.35% of students will have height within the range of 168-175cm. This percentage corresponds to 500 x 0.0535 = 26.75 which approximates to 27 students.

3.3 Continuous Probability Distribution

Continuous probability distribution is the distribution of continuous random variables and the best examples is the normal distribution where;

- The observations can fall anywhere within an interval.
- Pr (a x b) is needed.
- A graphical representation is the histogram and the limiting form will be a smooth curve.
- The total area under the smooth curve called the probability density function equals unit.
- The probability of a value between a and b is the area under the curve between a and b.

4.0 CONCLUSION

In this unit we have discussed the standard normal distribution as one with mean of zero and Standard deviation of one as well as the total area under the curve equal to unit. Instead of x along the X-Axis we have a transformation of x called the standard score Z. Thus the Z-score really tells us how many standard deviations from the mean a particular Z-score is. The importance of the Z distribution was explained to be that the probability of a value being anywhere between two points is equal to the area under the curve between those two points. It was also noted that another important use of Z derives is that we can also convert a sample mean (rather than just a single individual value) to a Z-score.

5.0 SUMMARY

In this unit we learnt:

- the concept of Standard Normal Distribution
- the properties of Standard Normal Distribution
- the importance of Standard Normal Distribution
- the difference and uses of Standard Normal Distribution and Z-Score.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define and explain the concept in Standard Normal Distribution.
2. Mention the properties of Standard Normal Distribution.
3. Discuss the importance of Standard Normal Distribution.
4. Explain the difference and uses of Standard Normal Distribution and Z-Score.

7.0 REFERENCES/FURTHER READING

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York: 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Dorling Kindersley Pvt. Ltd. Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications, Ltd. Malden: 3-15.

UNIT 3 BINOMIAL AND POISSON DISTRIBUTIONS

CONTENTS

1.0 INTRODUCTION

The Binomial distribution is relevant in cases where there are two outcomes to a trial, success of probability ọpô or failure ọqô and the trial is repeated ọnô times with each repetition being independent. While the Poisson distribution is a discrete probability distribution of a number of events occurring in a fixed period of time if these occur with known average rate and are independent of the time since the last event. Poisson distribution occurs when there are events which do not occur as outcomes of a definite number of trials but at random points of time and space.

2.0 OBJECTIVES

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

- define Binomial Distribution
- understand the concept of Binomial Distribution
- know how to calculate, interpret and apply Binomial Distribution
- define Poisson Distribution
- understand the concept of Poisson Distribution
- know how to calculate, interpret and apply Poisson Distribution

3.0 MAIN CONTENT

3.1 Binomial Distribution

The Binomial distribution is relevant in cases where there are two outcomes to a trial, success of probability ọpô or failure ọqô and the trial is repeated ọnô times with each repetition being independent. It is mathematically expressed as

$$Pr(r) = \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

- **Binomial application**

It was observed in a health facility that when tetanus affects newborn infants only 10% recover, what is the probability that;

i. Two such affected newborns will recover i.e. $r = 2$
ii. None such newborns will recover i.e. $r = 0$
iii. At least four such newborns will recover i.e. $r = 4$ or 5

In a random sample of 5 affected newborns?

Solution: Given $P = \frac{1}{10}$, $q = \frac{9}{10}$, $n = 5$, $r = 0, 1, 2, 3, 4, 5$.

- Probability for two to recover will be

$$Pr = \frac{5!}{2!(5-2)!} \times \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^{5-2}$$

$$Pr = \frac{120}{2 \times 6} \times \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^3$$

$$= \frac{10}{1} \times \frac{1}{100} \times \frac{729}{1000} = 0.073 \text{ or } 7.3\%$$

- Probability for none to recover will be

$$Pr = \frac{5!}{0!(5-0)!} \times \left(\frac{1}{10}\right)^0 \left(\frac{9}{10}\right)^{5-0}$$

$$Pr = \frac{5}{1} \times (1)\left(\frac{9}{10}\right)^5 = 0 \times 1 \times \frac{59049}{100,000} = 0.59 \text{ or } 59\%$$

- Probability for at least four will recover i.e. Pr(4)+Pr(5)

For four to recover $Pr(4) = \frac{5!}{4!(5-4)!} \times \left(\frac{1}{10}\right)^4 \left(\frac{9}{10}\right)^{5-4}$

$= 0.00005$

For five to recover $Pr(5) = \frac{5!}{5!(5-5)!} \times \left(\frac{1}{10}\right)^5 \left(\frac{9}{10}\right)^{5-5}$

$= 0.00001$

$Pr(4)+Pr(5) = 0.00005 + 0.00001$

$= 0.00006$ or 0.006%

We get Binomial distribution under the following experimental conditions;

- Each trial results in two mutually exclusive disjoint outcomes, termed as success and failure.
- The number of trials is finite.

- The trials are independent of each other.
- The probability of success ópô is constant for each trial.

- **Binomial Law of Probability**

In Biological Science this law can be applied when two children are born one after the other then the probability sequence will be any of the followings;

1st Issue	2nd Issue	Sequence	Probability	% chances
M		M, M	½ x ½ = ¼	25
M	F	M, F	½ x ½ = ¼	25
M		F, M	½ x ½ = ¼	25
F	F	F, F	½ x ½ = ¼	25

Fig. 20. Distribution of Childôs sex Probability.

The probability of the two issues being males is ¼ or 25%
Also the probability of the two issues being females is ¼ or 25%
While the probability of either male or female is ¼+¼ = ½ or 50%

3.2 Poisson distribution

The Poisson distribution is a discrete probability distribution of a number of events occurring in a fixed period of time if these occur with known average rate and are independent of the time since the last event. Poisson distribution occurs when there are events which do not occur as outcomes of a definite number of trials but at random points of time and space.

Poisson distribution is sometimes referred to as an approximation to the Binomial distribution when the probability of a success ópô is small and the number of trials ónô is large. It is mathematically expressed as;

$$Pr(r) = \frac{\mu^r e^{-\mu}}{r!}$$

Where e = 2.718,
r! = (r)(r-1)(r-2).....
μ = mean value

A random variable follows a Poisson distribution if it assumes only non-negative values. Poisson distribution is a limiting case of Binomial distribution under the following conditions;

a. ónô the number of trials indefinitely large $n \longrightarrow Ð$
b. ópô the constant probability of success for each trial and is definitely small, that is given as $p \longrightarrow 0$

Poisson distribution is useful in bacteriology for the estimation of live organism in a suspension. It is also used in epidemiology for counting rare events in a population.

- **Poisson Application**

The number of deaths from neonatal tetanus cases presenting in a clinic averages 4 per year. Assuming a Poisson distribution is appropriate, the probability that 6 deaths due to neonatal tetanus cases presenting at same clinic yearly will be;

$$\Pr(r) = \frac{\text{[illegible]}}{\text{[illegible]}} = \frac{\text{[illegible]}}{\text{[illegible]}} = \frac{\text{[illegible]}}{\text{[illegible]}} = \frac{\text{[illegible]}}{\text{[illegible]}} = 0.1 \text{ or } 10\%$$

4.0 CONCLUSION

In this unit we learnt that Binomial distribution is relevant in cases where there are two outcomes to a trial, success of probability ópô or failure óqô and the trial is repeated ónô times with each repetition being independent. While in Poisson distribution it is a discrete probability distribution of a number of events occurring in a fixed period of time if these occur with known average rate and are independent of the time since the last event. Poisson distribution occurs when there are events which do not occur as outcomes of a definite number of trials but at random points of time and space. Again Poisson distribution is sometimes referred to as an approximation to the Binomial distribution when the probability of a success ópô is small and the number of trials ónô is large.

5.0 SUMMARY

In this unit we have learnt:

- the define Binomial Distribution
- and understood the concept of Binomial Distribution
- how to calculate, interpret and apply Binomial Distribution
- the define Poisson Distribution
- and understood the concept of Poisson Distribution
- how to calculate, interpret and apply Poisson Distribution.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define Binomial Distribution.
2. Explain the concept in the application of Binomial Distribution.
3. Explain how to calculate, interpret and apply Binomial Distribution.
4. Define Poisson Distribution.

5. Explain the concept in the application of Poisson Distribution.
6. Explain how to calculate, interpret and apply Poisson Distribution.

7.0 REFERENCES/FURTHER READING

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

Richard, F. Morton and Richard, J. Hebel. (1979). *A study guide to Epidemiology and Biostatistics*. University Press. New York: 59-109.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Habason Nig. Limited. Kano: 32-34.

UNIT 4 SKEWED DISTRIBUTION

CONTENTS

1.0 INTRODUCTION

A distribution is said to be skewed if it is asymmetrical. It is a uni-modal frequency distribution and could be left or right skewed. Median is a preferred measure of central tendency to mean in a skewed distribution because the mean is unduly weighted by outliers that skewed the distribution. Skewed distribution is positive if mean is greater than median. Skewed distribution is negative if mean is less than median.

2.0 OBJECTIVES

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

- mention what skewed distribution is
- identify the various types of skewed distribution
- discuss their applications
- explain the implications of skewed population
- state the differences between normal population distribution and skewed population
- state the relationship of kurtosis and measures of central tendency.

3.0 MAIN CONTENT

3.1 Right skewed

The distribution is right skewed if its longer tail is extending towards the higher values of the variable. It is also known to be positively skewed.

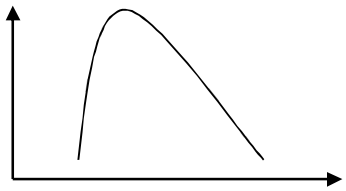


Fig. 23a. Right skewed distribution.

3.2 Left skewed

The distribution is left skewed if its longer tail extending towards the lower values of the variables. It is also known to be negatively skewed.

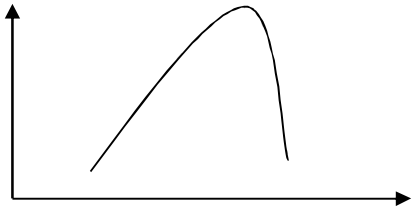


Fig. 23b. Left skewed distribution.

Mean-Mode ~ 3(Mean ï Median) for moderately skewed population
Skewness of a distribution can be measured by the following formulae;

- Skewness = Mean ï Median
- Skewness = Mean ï Mode
- Coefficient of Skewness = $\frac{(\square\square\square\square\square\square\square\square\square)}{\square}$ This is used for comparing two series.
- Skewness is positive if Mean > Mode or Mean > Median
- Skewness is negative if Mean < Mode or Mean < Median

3.3 Kurtosis

Kurtosis refers to the flatness of the curve.

- **Mesokurtic Curve:** Normal curve.
- **Platykurtic Curve:** Flatter than normal curve.
- **Leptokurtic Curve:** More peaked than normal curve.

4.0 CONCLUSION

In this unit, we discussed skewed distribution to be asymmetrical with unimodal frequency distribution which could be left or right skewed. When a population distribution is skewed Median is a preferred measure of central tendency to mean because the mean is unduly weighted by

outliers that skewed the distribution. Skewed distribution is positive if mean is greater than median and negative if mean is less than median.

5.0 SUMMARY

In this unit we have learnt that:

- a distribution is said to be skewed if it is asymmetrical.
- median is a preferred measure of central tendency to mean in a skewed distribution
- distribution is positive if mean is greater than median.
- skewed distribution is negative if mean is less than median.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define a skewed distribution.
2. Identify and explain the various types of skewed distribution.
3. Mention the applications of skewed population distribution.
4. State the implications of skewed population.
5. What are the relationships of kurtosis and measures of central tendency?

7.0 REFERENCES/FURTHER READING

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Habason Nig. Limited. Kano: 32-34.

Syvia, Wassertheil-Smoller. (1990). Biostatistics and epidemiology. A primer for health professionals. Springer-Verlag. New York: 119.

Wayne, W. D. (2006). *Biostatistics. A Foundation for Analysis in the Health Sciences*. 7th edition. John Wiley and Sons. New Delhi: 57-71.

MODULE 7 SAMPLING AND SAMPLING TECHNIQUES

UNIT 1 SAMPLING AND DEFINITION OF TERMS

CONTENTS

1.0 INTRODUCTION

Statistical inference relies on inductive projection of findings from the sample to the population which is based on the assumption that the sample is representative of the population from where it is drawn. This representativeness is ensured by random sampling in which the individual units in the population are selected by chance and at a given equal opportunity of being selected. A sample is therefore a subset of the larger population or parent population

2.0 OBJECTIVES

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

- define and understand the following terms; parent population, target population, sample statistic, sample size, sampling frame, sampling unit, sampling fraction, sample error and sample.
- discuss the concept of sampling
- state sources of bias in sampling.

3.0 MAIN CONTENT

3.1 Definition of Terms

- **Parent Population:** This is the group of individuals under study. This may be finite or infinite. More often than not practical problems make it difficult for the investigator to cover the total population. In such cases what he covers is called the survey population i.e. the population that is accessible to him now referred to as the sample.
- **Target population:** This is the aggregate of all the units about whom information is required e.g. the study of pregnant women in a community.
- **Sample Statistic:** This is the measures computed from the sample observation alone e.g. mean (□) and standard deviation (ů).
- **Sample Size.** This is the total number of units in the sample.
- **Sampling Frame.** This is the list containing all the units in the population e.g. the list of all pregnant women in the community. It may be a register or any kind of directory. Sampling frame is expedient in most probability samplings.
- **Sampling Unit.** This is the smallest unit in the selection process. This should be well defined in every survey e.g. a household or an under five in community survey of nutritional status or a pregnant woman studied in the community.
- **Sampling Fraction.** This is the proportion of the total population that is constituted by the sample. If the target population is 100 and the sample size to be studied is 20 the sampling fraction is $^{20}/_{100} = {}^{1}/_{5}$. The inverse $^{5}/_{1}$ i.e. 5 becomes the sampling interval in that every 5^{th} unit is selected in a series.
- **Sample Error:** This is an index of the precision of the estimate obtained from a sample.

3.2 Sample

This is a finite subset of individuals in a population selected for study. A sample should represent the population adequately as the purpose is to infer the characteristics of the population from the sample findings. Sampling is based on two principles viz.

1. **Elimination of bias** by ensuring that,
    - All units have equal chance of being selected.
    - Every unit has a known chance of being selected.
    - Probability sampling technique is used to select each unit.

2. **Obtaining High Precision**. Precision is the measure of the way in which repeated observations or outcomes conform to the others. It is also called the error margin. It is the quantity obtained by multiplying the reliability factor by the standard error of the mean.

3.3 Sources of Bias

1. **Design defect.** When the wrong study design is adopted the study will be defective as bias is being introduced from the on-set of the study.
2. **Observer error.** This may be inter observer error where biases are introduced by different persons whose observations or measurements are in variance. It may also be intra observer error where the same personôs observations or measurements on the same variable are in variance.
3. **Instrument error.** Biological measurements are also subject to variability. This variability may be inherent to the instrument, peculiar to environmental factors e.g. climates, variation within an individual, from one occasion to another and from one observer to another etc. Variation in instruments can introduce errors in results or outcomes. Therefore, in order to assess biological data we need statistical techniques that will help us cope with such variability.
4. **Communication problems.** Poor communication, no response, incomplete or misinformation may occur between observer and respondent.

3.4 Advantages of sampling

1. It reduces cost of study.
2. Not all persons in the population are studied.
3. Material used is less.
4. Demand on personnel is less.
5. It guarantees quick result as smaller population is studied.
6. Reduces time constraint.
7. Reduces error and enhances accuracy due to smaller population studied.

3.5 Limitations of sampling

1. Some people or units are excluded.
2. Sample mean may not be equal to population mean i.e. $\bar{x} \neq \mu$
3. It may be difficult sometimes to select a sample that is representative of the population.

4. In human population it is naturally easy to introduce bias (discrimination) in sampling.
5. Some surveys may not fit into sampling as everyone has to be interviewed, e.g. census.

4.0 CONCLUSION

In this Unit we explained that statistical inference relies on inductive projection of findings from the sample to the population which is based on the assumption that the sample is representative of the population from where it is drawn. This representativeness is ensured by random sampling in which the individual units in the population are selected by chance and a given equal opportunity of being selected. A sample is therefore a subset of the larger population or parent population. We also defined and explained certain terms used in population sampling viz. Parent population, Target population, Sample statistic, Sample size, Sampling frame, Sampling unit, Sampling fraction, Sample error and Sample. Ways of eliminating bias during sampling and enhancing precision were discussed while the advantages and limitation in sampling were explained.

5.0 SUMMARY

In this unit we have learnt:

- findings from the sample to the population is based on the assumption that the sample is representative of the population from where it is drawn.
- representativeness in sampling is ensured by random sampling
- the individual units in the population sample are selected by chance and a given equal opportunity of being selected.
- a sample is a subset of the larger population or parent population.
- the definitions of Parent population, Target population, Sample statistic, Sample size, Sampling frame, Sampling unit, Sampling fraction, Sample error and Sample.
- it is important to eliminate bias during sampling through using probability sampling techniques
- the choice of sampling techniques that enhance precision is important

6.0 TUTOR-MARKED ASSIGNMENT

1. What does statistical inference rely on?
2. Finding from the sample to the population is based on what?
3. How do we ensure representativeness in sampling?

4. What are the two main factors considered in selecting sample units that makes the process scientific and probability sampling?
5. What is the relationship of a sample to the larger population or parent population?
6. Define Parent population, Target population, Sample statistic, Sample size, Sampling frame, Sampling unit, Sampling fraction, Sample error and Sample.
7. What is sampling bias and how is it eliminated?
8. How do we enhance precision during sampling?

7.0 REFERENCES/FURTHER READING

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York: 1-8.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications Ltd. Malden: 3-15.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Edition. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

UNIT 2 NON-PROBABILITY SAMPLING TECHNIQUES

CONTENTS

1.0 INTRODUCTION

Non-probability sampling technique is a non-scientific method of selecting sampling units from the population as some elements of bias are introduced in the process. All these sampling techniques being subjective give no basis to determine whether the chosen population truly represents or retains the true characters of the parent population. Under normal circumstances we do not use Non-probability sampling techniques because we cannot use the mathematics of the probability to analyse the data generated through the non-probability sampling technique. In general we cannot count on a non-probability sampling approach to produce representative samples and infer our findings on the general population.

2.0 OBJECTIVES

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

- define non-probability sampling techniques
- provide reasons why non-probability sampling techniques are not normally used
- explain the drawbacks of non-probability sampling techniques
- list the non-probability sampling techniques
- describe the procedures in the method of non-sampling techniques

- discuss situations where non-sampling techniques may be applicable.

3.0 MAIN CONTENT

3.1 Quota Sampling

In Quota sampling the population is first segmented into mutually exclusive sub-groups just as in the case of stratified sampling technique. Then from each sub-group the sampling units are selected into sample on the basis of pre-specified characteristics so that the total sample has the same distribution of characteristics assumed to exist in the population being studied. Therefore, in order to maintain the representative characteristics of the sample, fixed number or quota or proportion of the sub groups which constitute the population is included in the sample e.g. allocated specific proportion to each village in a local Government Area.

3.2 Convenience Sampling

Convenience sampling is sometimes referred to as grab or opportunity, accidental or haphazard sampling. The units are selected based on the researcherôs convenience. Units happen to be available and may have been selected for some other purpose. A researcher may decide to choose a particular area because it is feasible for the study without logistic constraints. The researcher using such a sample cannot scientifically make generalizations about the study population from the sample because it will not be representative enough.

3.3 Conformance Sampling

Selection of units is based on expert opinion or on judgement of the parent population.

3.4 Compliance Sampling

Units are available to the investigator by permission only. Therefore the researcher is restricted to sampling units made available to him.

3.5 Purposive or Judgemental Sampling

In purposive sampling the researcher chooses the sample based on who or location which he thinks is appropriate for the study without applying any scientific method of selection. This is done primarily when there is a limited number of people that have expertise in the area being researched or the study area is selected based on personal justifications.

3.6 Panel Sampling

Panel sampling is a sampling method of first selecting a group of participants through a random sampling method and then asking the same group the same information again several times over a period of time. Therefore each participant is given the same survey or interview at two or more time points with each period of data collection called a ówaveô. This sampling method is often applied to large scale or nation-wide studies in order to gauge changes in the population with regard to any number of variables from chronic illness to job stress to weekly food expenditures. Panel sampling can also be used to inform researchers about within-person health changes due to age or help explain changes in continuous dependent variables such as spousal interaction.

3.7 Event Sampling

Event sampling method is a sampling method that allows researchers to study on-going experiences and events that vary across and within days in its naturally occurring environment. Participants are asked to record their experiences and perceptions in a paper or electronically. Due to the frequent sampling of the events the researcher is able to measure the typology of the activities and detect the temporal and dynamic fluctuations occurring. It addresses the shortcoming of cross-sectional research as researchers can now detect intra-individual variances across time. The various types of Event sampling are;

- **Signal Contingent.** In Signal contingent method there is a random beeping that notifies the participant to record data thereby minimizing recall bias.
- **Event Contingent.** In Event contingent method the participant records data when events occur. Event Sampling Method may be considered invasive and intrusive by participants resulting in unwillingness to participate or cooperate in the data collection. There may also be self-selection bias.
- **Interval contingent.** In Interval contingent the participant records data according to the passing of certain period of time.

3.8 Snowball Sampling

A snowball sample is a non-probability sampling technique that is appropriate for use in a research when the subjects are difficult to locate in the target population. Therefore the researcher has no option but to collect data on the few members of the target population he or she can locate. Thereafter the researcher asks those individuals he was able to

reach to provide information needed to locate other members of that population whom they know how to locate.

Snowball sampling techniques will hardly and likely give a representative sample of the target population, but there are times when it may be the best or only method feasible. An example of Snowball is when a researcher is carrying out a study on the homeless, he is unlikely to produce a sampling frame or find a list of all the homeless people in the city. However, if the researcher identifies one or two homeless individuals that are willing to participate in the study, it is likely that they know other homeless individuals in their area and can help in locating them. The same goes for underground sub-cultures, or any population that might want to keep their identity hidden, such as undocumented immigrants, ex-convicts, those living with HIV/AIDS co-educators, etc.

Because snowball sampling is hardly representative of the larger study population, it is primarily used for exploratory purposes. That is, the researcher is ñfeeling outò a topic or population to study further in-depth at a later time.

4.0 CONCLUSION

In this unit we explained how Non-probability sampling technique is a non-scientific method of selecting sampling units from the population. This is so because some elements of bias are introduced in the process of selecting the sampling units. All the non-probability sampling techniques have elements of subjectivity with no basis in determining whether the chosen population truly represents or retains the true characters of the parent population. For these obvious limitations and under an ideal situation Non-probability sampling techniques are not used because data collected through such process cannot be subjected to the mathematics of probability analysis. Therefore it was explained that in general we cannot count on a non-probability sampling approach to produce representative samples from a given population let alone inferring our findings on the general population.

5.0 SUMMARY

In this unit we:

- defined Non-probability sampling technique
- gave the basis for which it is not a scientific process
- stated that under normal circumstances we do not use Non-probability sampling techniques

- stated that the reason why non-probability sampling technique is not usually used is because we cannot use the mathematics of the probability to analyse the data generated
- we explained that in general we cannot count on a non-probability sampling approach to produce representative samples and infer our findings on the general population.
- defined and explained the following non-probability sampling techniques; Quota Sampling, Convenience Sampling, Conformance Sampling, Compliance Sampling, Purposive or Judgemental Sampling, Panel Sampling, Event Sampling, Snowball Sampling.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define non-probability sampling techniques.
2. Provide reasons why non-probability sampling techniques are not normally used.
3. Explain the drawbacks of non-probability sampling techniques.
4. List the non-probability sampling techniques.
5. Describe the procedures in the method of non-sampling techniques.

7.0 REFERENCES/FURTHER READING

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

UNIT 3 PROBABILITY SAMPLING TECHNIQUE

CONTENTS

1.0 INTRODUCTION

Probability sampling techniques apply scientific sampling techniques in selecting every sampling unit in the population i.e. each unit has equal chance which is greater than zero of being selected. Probability sampling technique reduces bias and also enhances precision. This technique can further be classified into random sampling technique and non-random sampling technique. These are probability sampling techniques that apply scientific sampling techniques in selecting every sampling unit in the population i.e. each unit has equal chance which is greater than zero of being selected in the sample and this probability can be accurately determined. Being a scientific method it is based on certain basic postulations of methods viz.

- Rely on empirical evidence.
- Utilize prove and relevant concepts.
- Result into probabilistic predictions.
- Are committed to only objective considerations.
- Presupposes ethical neutrality.
- Formulating most general axioms or what can be termed as scientific theories.
- Are made known to all concerned for critical scrutiny.
- Findings and conclusions can be tested and replicated.

Probability sampling technique reduces bias and also enhances precision. This technique can further be classified into random sampling technique and non-random sampling technique.

2.0 OBJECTIVES

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

- define and explain Probability sampling techniques and how it applies scientific techniques in selecting sampling units in the population
- discuss the concept in probability sampling techniques
- explain certain basic postulations in probability sampling technique
- explain how Probability sampling technique reduces bias and also enhances precision.
- classification probability sampling techniques into random sampling technique and non-random sampling techniques.
- state the two basic features of probability sampling techniques that make it a scientific process
- list the basic postulations considered in probability sampling technique.

3.0 MAIN CONTENT

Probability sampling technique can be classified into Random and Non-random probability sampling Techniques

3.1 The Random probability sampling technique

In this technique ALL sampling units are selected at random giving each equal opportunity of being selected. Examples are;

Simple Random Sampling: All units have equal chance of being selected and this probability is independent of the previous drawing. The various method used are;

- Use of table of random numbers.
- Use of computer
- Balloting can be; Sampling with replacement for finite population or Sampling without replacement for infinite population.

Advantages

- Has high precision for homogeneous population.
- Eliminates bias
- It is an equitable method of selection.
- Estimates are easy to calculate.

Disadvantages

- Requires a sampling frame which may not be feasible in an infinite population.
- Minority sub-groups of interest in study population may not be fully represented.
- May not produce a good representation sample in heterogeneous population.
- Cumbersome in large population.
- The structural peculiarity of the population is not utilized.

Stratified Sampling: Before the selection the entire population is divided into homogeneous groups usually called Strata according to some relevant characteristics of the population structure. The sampling units are thereby randomly selected from each Stratum using simple random sampling. Allocation of the number of sample units in different strata when the overall sample size is fixed in advance may be done in the following ways;

- Equal allocation. Equal numbers of units are selected from the various groups or strata in the population.
- Proportional allocation. The numbers of units selected from each stratum is proportional to the stratum.
- Optimum allocation. The number of units to be drawn from each stratum is proportional to the standard deviation of the stratum.

Advantages

- Sub-groups in the population are adequately represented.
- Has a high precision of estimates in heterogeneous population.
- Provision of reasonably accurate estimate for subgroups of interest.

Disadvantage

- Construction of sampling frame may be cumbersome.

Cluster Sampling: When natural groupings are evident in the population, rather than single subjects being selected from the population the subjects are selected from identified groups or clusters using appropriate random sampling technique. This is a case of the multistage sampling whereby all units are studied at the second stage.

- Single-stage Cluster Sampling: All the units in the selected cluster are studied.
- Two-stage Cluster Sampling: An appropriate random sampling technique is used to select sampling units from each selected cluster.

- Area or Geographical Cluster sampling: In geographically dispersed population simple random sampling can be achieved by treating several respondents within a local area as a cluster.

Advantages

- Provides easier survey logistics e.g. in travelling and administrative costs.
- Reduces cost of field work as selection takes place only once.
- Does not require a sampling frame and advantageous in rural area.
- Provides judicious use of research resources by avoiding duplication.

Disadvantages

- Important group may be missed out during the selection.
- May not be representative of whole population.
- Precision is low because many first stage units are excluded.
- Requires more complicated data analysis.
- May waste resources if units in a cluster are similar or same.

Multistage Sampling: This is sampling done in stages. More than one sampling technique is applied in selecting the sampling unit from the total population e.g. Stratified sampling technique can be used to select groups of heterogeneous population and then simple random sampling technique used in selecting the sampling units for study e.g. a study on undergraduates, select 20 departments out of the total departments at random (balloting) at the first stage. Second stage, stratify the students by level and select levels as clusters. Third stage, stratify the students by their sex and from sex stratum proportionally select sampling units.

Advantages of Multistage sampling technique

- Has high precision.
- Preparing sampling frame is less cumbersome.
- All groups are well represented in the final selection.

Disadvantage of Multistage sampling technique

- Consumes more money and time.

Multiphase Sampling: In multiphase sampling a part of the information is collected from the whole sample and part from the subsample. With each successive sampling the size becomes smaller. Each sampling unit is studied more than once e.g. screening test at the initial stage before confirmatory test. In a tuberculosis survey, the first phase will have everybody do a Mantoux test. All those testing positive will be subjected to the second phase of Chest X-ray and the third face of sputum test

those with positive Chest X-ray. It is used to obtain additional information. It provides the examination of units in more details.

3.2 Non-random probability Sampling Technique

Systematic or quasi-random Sampling Technique: In Systematic sampling technique the sampling units are selected after certain pre-determined sampling intervals. In order to make it a purely random sampling the first unit should be drawn at random while subsequent units follow the interval. The sampling interval is the ratio of total population and the sample size.

Advantages

- Sampling units are easier to draw without mistakes.
- In some surveys may not require a sampling frame where arrivals are at random and every arrival at a known interval e.g. outpatient clinic.
- The sample is usually spread evenly over the entire population.
- It has a high precision in an evenly spread population.

Disadvantages

- Unreliable or biased in periodic arrangements as hidden periodicity in population may coincide with that of selection.
- Precision of the estimate difficult to measure especially in one survey.
- Sampling frame may be difficult to produce in large population.

3.3 Generating Random Numbers

Microsoft Excel has a function to produce random numbers. The function is simply given as; **=RAND( )**. Type that into a cell and it will produce a random number in that cell. Copy the formula throughout a selection of cells and it will produce random numbers between 0 and 1. If you would like to modify the formula, you can obtain whatever range you wish. For example if you want random numbers from **1** to **250**, you could enter the following formula: **=INT[250*RAND( )]+1.**

The **INT** eliminates the digits after the decimal, the **250*** creates the range to be covered, and the **+1** sets the lowest number in the range. See Appendix II for table of random numbers.

3.4 Definition of Terms

- **Sampling Error:** Sampling error is a consequence of the non-totality of the information contained in a sample leading to some error in the estimates of the population parameters. This type of

error is minimized by large sample size and using appropriate sample design with high precision.

- **Sampling Distribution:** The sampling distribution of a statistic is the distribution obtained by computing the statistic for all possible samples of a predetermined size which can be drawn from a given population. Sampling distribution is approximately normal when the distribution of parent population is normal or in any form provided $n \geq 30$.
- **Standard Error:** The Standard Error of a sample is the Standard deviation of the distribution of a statistic. It gives an indication of the adequacy of the statistic as an estimator of the parameter. It is a function of the variability of the factor and sample size.
- **Precision.** Precision refers to the extent to which repeated observations conform or agree to themselves. It is a measure of reliability of the sample statistic as an estimate of the population parameter.
- **Accuracy:** Accuracy refers to the closeness of an observed value of a quantity to its true value.
- **Degrees of Freedom:** In statistics, the number of **degrees of freedom** is the number of values in the final calculation of a statistics that are free to vary. This is the number of independent ways a dynamic system can move without violating any constraint imposed on it. The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom. It is an estimate of a parameter which is equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself, which, in sample variance, is one, since the sample mean is the only intermediate step. In the calculation of variance the denominator is (n-1) and called the number of degrees of freedom of the variance. This number is (n-1) rather than (n), since only (n-1) of the deviations (X-□) are independent from each other. The last one can always be calculated from the others because all (n) of them must add up to zero. Degree of freedom is commonly applied in relation to tests statistics like the Chi-Square and t-tests when determining the significance of the test statistics and the validity of the Null hypothesis.

4.0 CONCLUSION

In this unit we have discussed Probability sampling techniques as an application of scientific sampling techniques in selecting every sampling unit in the population whereby every unit has equal chance greater than zero of being selected. It was further stressed that Probability sampling technique reduces bias and also enhances precision. Broadly speaking

we identified two distinct classification of the technique in the context of random sampling in selecting the first sampling unit. These are probability sampling techniques that apply scientific sampling techniques in selecting every sampling unit in the population i.e. each unit has equal chance and this probability can be accurately determined. Being a scientific method it relies on empirical evidence, utilizes prove and relevant concepts, result into probabilistic predictions, are committed to only objective considerations, presupposes ethical neutrality, formulates most general axioms or what can be termed as scientific theories, are made known to all concerned for critical scrutiny, and the findings and conclusions can be tested and replicated.

5.0 SUMMARY

In this unit we have learnt:

- how to define and explain Probability sampling techniques and how it applies scientific techniques in selecting sampling units in the population
- and understood the concepts in probability sampling techniques
- how Probability sampling technique reduces bias and also enhances precision.
- the classification of probability sampling techniques into random sampling technique and non-random sampling techniques based on selected the first sampling unit.
- the two basic features of probability sampling techniques that make it a scientific process
- the basic postulations considered in probability sampling technique.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define and explain Probability sampling techniques.
2. How does Probability Sampling Technique apply scientific techniques in selecting sampling units from the population?
3. Explain the concepts applied in probability sampling techniques.
4. What are the basic postulations in probability sampling technique?
5. Specify how Probability sampling technique reduces bias and also enhances precision.
6. Classify probability sampling techniques into random probability sampling technique and non-random probability sampling techniques.
7. What are the two basic features of probability sampling techniques that make it a scientific process?

7.0 REFERENCES/FURTHER READING

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York: 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Dorling Kindersley Pvt. Ltd.Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications Ltd. Malden: 3-15.

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

MODULE 8 INFERENTIAL BIOSTATISTICS

UNIT 1 CONCEPT AND CLASSIFICATION

CONTENTS

1.0 INTRODUCTION

In this unit we will discuss the concepts in inferential statistics and the two instruments of drawing inference on population statistic. Descriptive statistics aggregates population data by compartmentalizing it through the instrumentalities of tables, diagrams and numerals but falls short of drawing conclusion on observed features or parameters of the population. However, hypothetical statements can be formulated from such findings and it is through inferential statistics that such hypothetical statements can be tested. Inference can only be drawn on the population about population statistics after such has been subjected to statistical tests and probability of committing statistical error determined.

2.0 OBJECTIVES

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

- define and explain inferential statistics
- understand and explain the concepts in inferential statistics
- define and differentiate between inductive and deductive inference
- explain the applications of inductive and deductive inferences

3.0 MAIN CONTENT

3.1 Concept and classification

It was after several centuries of the practice of dogma that despised scientific reasoning that Francis Bacon advocated direct observation of phenomena, arriving at conclusions or generalizations through the evidence of many independent individual observations. This is an inductive process which deductive reasoning lacks. However, the deductive method of Aristotle and the inductive method of Bacon were fully integrated in the work of Charles Darwin in the nineteenth century.
In science we impose logic on observations we make. The logic of scientific reasoning is therefore entrenched in uncovering the truth in an investigation. In order to achieve this we apply two basic tools at our disposal which vary markedly from person to person. The tools are as follows; Our senses through which we experience the world around us and make kin observations; Our ability to reason which enables us make logical inferences on what we have observed and experienced.

We necessarily need our senses and the ability to reason as neither can on its own create a theory without the other. The scientific approach is, we make inductive inferences to generalize from many observations, make creative leaps of the imagination to infer explanations and construct theories, and use deductive inferences to test those theories. There are two kinds of logics in scientific inquiry to uncover the truth viz.

3.2 Deductive inference

In deductive inference we hold a theory and based on it we make a prediction of its consequences or what the observations should be. We move from the general i.e. from theory to the specific.

Inductive inference. In inductive inference we go from the specific to the general. We make many observations, design a pattern, make a generalization and infer an explanation. We move from the specific i.e. from observations to the general. Epidemiologists have generally been taught to use inductive inference. Science requires both deductive and inductive inferences as a particular point of view that provides a framework for observations which will lead to a theory. This will predict new observations which modify the theory leading to new predicted observations that will further modify the theory and lead to new predicted observations etc, etc. towards discovering the illusive truth. This can be illustrated with the diagram of deductive and inductive inference as shown in Fig. 24.

3.3 Inductive and deductive inference diagram

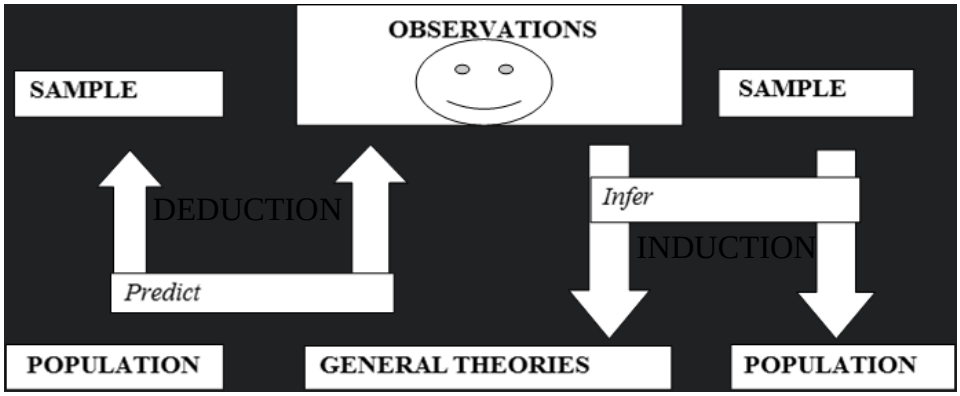


Fig. 24. Inductive and Deductive Statistics

4.0 CONCLUSION

In this unit we have been able to define and explain inferential statistics and the concepts in its applications. We have also been able to define, explain and differentiate between inductive and deductive inference as being used in inferential statistics. Also defined and discussed are significance tests, significance level, confidence level and confidence interval. The applications and interpretations of their results were also discussed. It was emphasized that it is not enough to say that there is a difference or association but that such difference be subjected to test of significance to find out whether the difference can be considered as a true difference or otherwise as the case may be. We can only conclude on such findings by ruling out a chance finding. This is accomplished by applying an appropriate significance test.

5.0 SUMMARY

In the unit we have learnt;

- that the logic of scientific reasoning is entrenched in uncovering the truth in an investigation.
- that we use our senses to make observations and ability to reason which enables us make logical inferences on what we have observed and experienced.
- that we necessarily need our senses and the ability to reason as neither can on its own create a theory without the other.
- that the scientific approach is, we make inductive inferences to generalize from many observations
- how to use deductive inferences to test those theories.

- there are two kinds of logics in scientific inquiry to uncover the truth viz: deductive and inductive inferences
- that in deductive inference we hold a theory and based on it we make a prediction of its consequences
- that in deductive inference we move from the general i.e. from theory to the specific.
- that in inductive inference we go from the specific to the general.
- that in science both deductive and inductive inferences will lead to a theory.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define and explain inferential statistics
2. Explain the concepts in inferential statistics
3. Define and differentiate between inductive and deductive inference
4. Define and explain significance test.

7.0 REFERENCES/FURTHER READING

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer-Verlag. 15.

Hannagan, T. (1997). *Mastering Statistics.* 3rd edition. NY: Palgrave Master Series. 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Dorling Kindersley Pvt. Ltd. Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications Ltd. Malden: 3-15.

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

UNIT 2 HYPOTHESIS TESTING

CONTENTS

1.0 INTRODUCTION

Hypothesis is a hypothetical statement about a population subject for verification based on information available from its subset (sample). It is a procedure whereby the truth of a Null hypothesis is investigated or otherwise by examining the value of the test statistic computed from a sample.

2.0 OBJECTIVES

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

- define and explain hypothesis
- define and explain null hypothesis
- define and explain alternative hypothesis
- explain and state the stages of Null hypothesis testing.

3.0 MAIN CONTENT

3.1 Null Hypothesis

Step 1: State the Null Hypothesis: Null Hypothesis (H_o) which states that there is no difference in the values of parameters being compared in the population. If any difference exists it should be due to chance or error or otherwise it is so small that it does not matter i.e.

H_o: $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$

Step 2: State the Alternative Hypothesis; H_A

Step 3: Plan, design and collect data using appropriate random sampling technique from the study population.

Step 4: Carry out descriptive statistics to have idea the distribution of the data to guide you in selection of Test statistic. This will provide explicit knowledge about the nature of population about which the hypotheses are set-up.

Step 5: Choose an appropriate Test Statistic which will reflect the probability of H_o and H_A.

Step 6: Determine the degree of freedom for appropriate test statistic where applicable.

Step 7: Compute the P-value

Step 8: Determine the critical region

Step 9: Compare computed value and critical value

Step 10: Decide on the Null Hypothesis and draw conclusion. On the basis of test Statistic reject Null hypothesis if the observed sample value falls outside the critical region and vice versa.

3.2 Alternative Hypothesis

Alternative Hypothesis (H_A) states that there is a difference in the values of parameters being compared in the population. This difference is not due to chance or sampling error and expressed thus;

H_A: $\mu_1 \neq \mu_2$ or $\mu_1 - \mu_2 \neq 0$

The H_A could be two-tail i.e. H_A : $\mu_1 \neq \mu_2$ (its direction is unknown).

But for one tail test the direction is known i.e. $\mu_1 < \mu_2$ for left tail and $\mu_1 > \mu_2$ for right tail test.

The distinction between one and two tail tests is relevant to the choice of critical region.

4.0 CONCLUSION

In this unit we were able to define and explain hypothesis which is a hypothetical statement about a population subject for verification based on information available from sample. It was explained that it is a procedure whereby the truth of a Null hypothesis is investigated by subjecting the test statistic computed from the sample to statistical test of significance. It was explained that there are two types of hypotheses viz. Null Hypothesis and Alternative Hypothesis. While the former is a

hypothesis of no difference the later says there is a difference which cannot be attributed to chance.

5.0 SUMMARY

In this unit we have learnt:

- the definition and application of hypothesis
- the definition and application of Null Hypothesis
- the definition and application of Alternation Hypothesis
- the difference between Null and Alternative Hypothesis
- the various stages in hypothesis testing.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define and explain hypothetical statement.
2. Define and explain the application of Null Hypothesis.
3. Define and explain the application of Alternation Hypothesis.
4. Differentiate between Null and Alternative Hypotheses.

7.0 REFERENCES/FURTHER READING

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York: 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Indore: Dorling Kindersley Pvt. Ltd. Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications, Ltd. Malden: 3-15.

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

UNIT 3 TESTS OF SIGNIFICANCE

CONTENTS

1.0 INTRODUCTION

Tests of significance are mathematical methods through which the probability or relative frequency of an observed difference that may have occurred by chance is detected. This difference may be found in a test of proportion or test of means between sample and population from where the sample was drawn or between two independent samples drawn from different populations or same population groups. This test of significance will lead to drawing inference or conclusion on the target population from where the studied sample was drawn. It is one of the methods of drawing conclusions or decision-making in science, the other being estimation method through construction of confidence intervals.

2.0 OBJECTIVES

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

- discuss and explain significance tests
- define and explain significance level
- define and explain confidence level
- define and explain confidence interval
- state the various test statistics used in test of significance

- choose an appropriate test of significance depending on the type of data collected and the objective of the study.
- decide the significance test to use by considering if the samples are paired or unpaired
- state whether your variables are Nominal (small or large sample), Ordinal or Numerical for an appropriate significance test.
- define critical region, critical level and critical value in test of significance
- define, explain and differentiate between Type 1 and Type 11 errors.

3.0 MAIN CONTENT

3.1 Test of Significance

Test of significance is a mathematical method applied to find out any difference occurring by chance. Choosing an appropriate significance test depends on the type of data collected and the objective of the study. In deciding the significance test to use we have to consider if the samples are paired or unpaired. Within each of these groups we also consider whether they are Nominal (small or large sample), Ordinal or Numerical.

3.2 Significance Tests

Significance test also known as statistical test is a range of tests used to analyse generated data under various conditions. Therefore, care has to be taken to use the correct significance test. Significance tests enable us decide if the difference observed between the sample statistic and hypothetical population parameter or between two independent sample statistics is significant or should be attributed to chance or sampling error.

A significant test therefore estimates the likelihood that an observed study result such as the difference between two or more groups is due to chance. Therefore, a significance test validates the finding of a study as to whether the same can be generalised on the total population from where the sample was drawn. There are different significance tests that have been developed for different sets of data. The likelihood or probability of observing a result by chance is usually expressed as a p-value which is expressed as a proportion. A probability of 5% corresponds to a p-value of 0.05. Therefore, a difference or an association is considered significant if $p<0.05$. This means that if the hypothesis stating that there is no difference between the groups compared is true, we would observe a difference in our data only 5 times or less in every 100 samples examined.

Suppose in a community survey on blood pressure, 40% of the women were hypertensive as against 35% of the men in the study sample. The question will be;

- Is the observed difference of 5% a true difference which can be inferred on the whole population (i.e. also exist in the total population from where the sample was drawn).
- Is the observed difference of 5% due to chance and does not reflect the reality in the total population i.e. due to sampling variation or bias from the sampling design.

It is also necessary to test the difference to find out whether the difference can be considered as a true difference or otherwise as the case may be in findings. We can only conclude on such findings by ruling out a chance finding. This is accomplished by applying a significance test.

3.3 Definition of Terms

- **Significant value**. This is the same as **Critical value** and it is the test statistic that separates the critical region (which is the rejection region) and the acceptance region. Significant value depends on;
  - The level of significance used in the test.
  - The Alternative Hypothesis i.e. whether it is one-tail or two-tail test.
- **Confidence limits.** Confidence limits are the outer boundaries which we calculate and about which we can say, we are 95% confident that these boundaries or limits include the true population mean. The interval between these limits is called the confidence interval.
- **Confidence level:** This is the probability that the value of a parameter falls within a specified range of values.
- **Confidence Interval.** This is a range of likely values for an unknown population parameter at a given confidence level.
  - Conventionally 95% confidence level is commonly chosen.
  - It tells us how much an exposure affects the subjects.
  - It conveys information on the magnitude of the differences between variables measured or compared.
  - Detects significant and non-statistically significant effects.

The tables 10-13 below match tests statistics with appropriate data and variable for the right tests of significance.

Table 10. Choosing significance test for differences between groups.

S/N	Variable	Data	Tests	
			Paired samples	**Unpaired samples**
1.	Nominal variables	Small sample size	Sign test	Fisherôs exact
		Large sample size	McNemarôs Chi-square test	Chi- square test
2.	Ordinal variables	Two groups	Wilcoxon signed-rank test	Wilcoxon two-sample test or Mann-Whitney U-test.
		More than two groups	Friedman 2-way analysis of variance	Kruskal-Wallis 1-way analysis of variance
3.	Numerical variables	Two groups	Paired test	t-test
		More than two groups		F-test

Table 11. Choosing Significance test when measuring association between variables.

S/N	Variable	Test	Also calculate
1.	Nominal variable	Chi-square if sample is large enough	Odds ratio or estimate relative risk
2.	Ordinal Numerical without linear relationship	Spearmanôs rho (r) Kendallôs Tau-b (Ú)	Significance
3.	Numerical with linear relationship	Pearsonôs correlation coefficient (Ɉ)	Significance

Table 12. Methods for comparing two samples

S/N	Type of data	Size of sample	Test
1.	Interval	Large >30 each sample. Small <30 each sample with nominal distribution. Small <30 each sample, non-nominal distribution.	Nominal distribution for mean. T distribution for means. Mann-Whitney U-test
2.	Ordinal	Any	Mann-Whitney U-test
3.	Nominal ordered	Large, most expected frequencies >5	Chi-square test.

4.	Nominal not ordered	Small, most expected frequencies >5. Small, more than 20% expected frequencies <5	Chi-square test Reduce number of categories by combining or excluding as appropriate.
5.	Dichotomous	Large, all expected frequencies >5. Small, at least one expected frequency <5.	Confidence interval for proportions, Chi-square test. Chi-square test with Yates correction, Fisherôs exact test.

Table 13. Methods for differences in one or paired samples

S/N	Type of data	Size of sample	Test
1.	Interval	Large >100 Small <100 normal differences Small < non-normal differences	Nominal distribution. Paired t test. Wilcoxon matched pairs test
2.	Ordinal	Any	Sign test
3.	Nominal ordered	Any	Sign test.
4.	Dichotomous	Any	McNemarôs test

Table 14. Methods for relationships between variables

	Interval normal	Interval non-normal	Ordinal	Nominal ordered	Nominal	Dichotomous
Interval normal	Regression Correlation	Regression Rank Correlation	Rank Correlation	Rank Correlation	ANOVA	t-test Normal test
Interval non-normal		Rank Correlation	Rank Correlation	Rank Correlation	ANOVA	Mann Whitney U test
Ordered			Rank Correlation	Rank Correlation	ANOVA	Mann Whitney U test
Nominal ordered				Chi-square Test for trend	Chi-square test	Chi-square test for trend
Nominal					Chi-square test	Chi-square test
Dichotomous						Chi-square test Fisherôs exact test

3.4 Critical region

Critical region is a set of values of the test statistics leading to rejection of Null hypothesis. When the value of the test statistics computed for the sample lies within the range of values specified as the critical region then the Null hypothesis is rejected. The location of the critical region is dependent on the sampling distribution of the test statistic and the significance level specification.

3.5 One-tail test

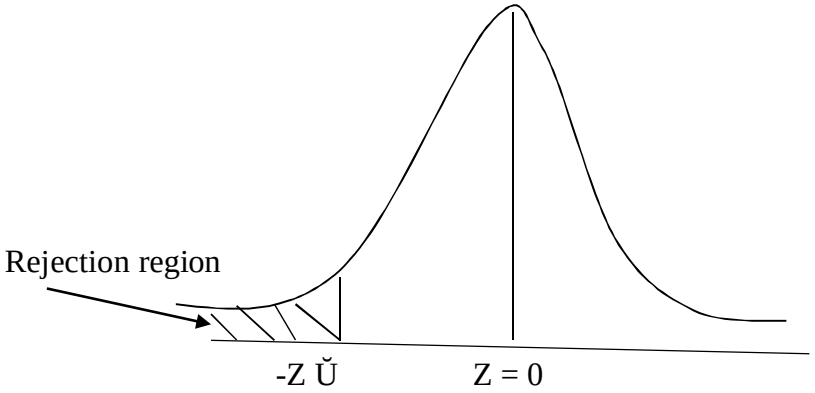


Fig. 25a. Left tail test distribution

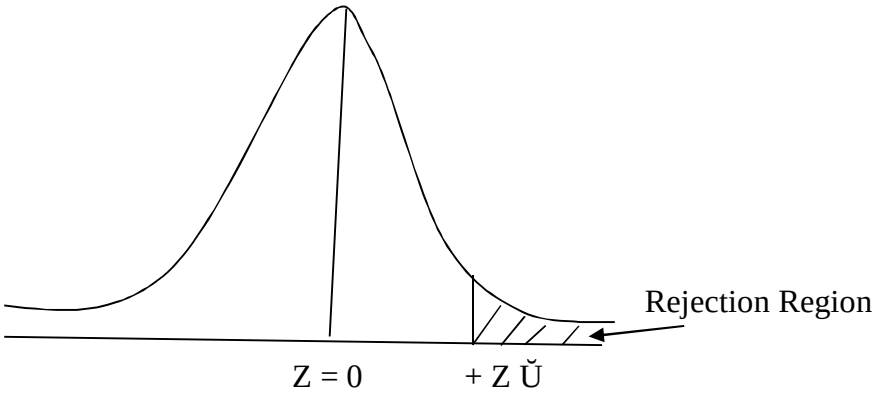


Fig. 25b. Right tail test distribution

3.6 Two-tail test

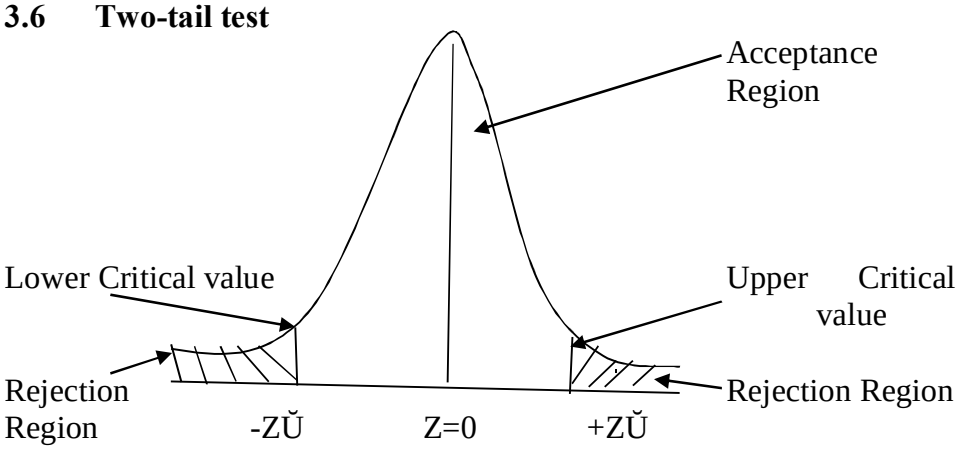


Fig. 25c. Two tail test distribution

Suppose that the critical value of the test statistics at a level of significance Ŭ for a two tail test is given by ZŬ is such that the area between the left of ï ZŬis $\frac{\square}{\square}$ and to the right of ZŬis also $\frac{\square}{\square}$,thus the total area Ŭis divided into two equal parts.

Table 15. Critical values (zŬ) of z

Critical Values (Zα)	**Significance Level**		
	1%	**5%**	**10%**
Two tail test	˚ZŬ˚ = 2.58	˚ZŬ˚ = 1.96	˚ZŬ˚ = 1.64
Right tail test	ZŬ= 2.33	ZŬ= 1.64	ZŬ= 1.28
Left tail test	-ZŬ= 2.33	-ZŬ= 1.64	-ZŬ= 1.28

3.7 Sample size consideration

When the sample is large usually taken as n>30 the population is assumed to closely approximate normal distribution therefore we can apply normal test statistic which is based on fundamental properties of normal probability curve. The following therefore applies;

- If the test statistic Z>3 $H_0$is therefore rejected.
- If the test statistic Z < 3 we further test at lower levels of significance e.g. 1%.

- For a two tail test, if the test statistic Z > 1.96, H_0 is therefore rejected at 5% level of significance.
- For a two tail test, if the test statistic Z > 2.58, H_0 is therefore rejected at 1% level of significance.

3.8 Type-1 error (α)

False positive is usually set at 5% i.e. 0.05 and called the level of significance. Type 1 error is therefore;

- the probability of rejecting Null Hypothesis when it is true.
- theŬwhich is the probability of Type-1 error known as the level of significance.
- Occurs when you discover a difference between two variables when indeed there is no such difference in the populations.
- Occurs when you establish an association where in reality there is no association between the variables.

3.9 Type-11 error (β)

False negative is usually set at 20% i.e. 0.2%. 1- ɓ is called the power of the test (1-0.20=0.80).

- Is the probability of accepting Null Hypothesis when it is false.
- Failure to reject Null Hypothesis when in reality it is false.
- Occurs when you found no association when in reality there is association between the variables in the populations.

(1- ɓ) is the power of test of Null Hypothesis against Alternative Hypothesis. Wrongly rejecting a Null hypothesis seems to be more serious error than wrongly accepting it. Since type-1 error is considered more serious than Type-11 error the usual practice is to control Type-1 error at a predetermined level Ŭand choose a test which minimizes ɓ.

3.10 Consequences of Type I and Type II errors

The relative seriousness of Type I and II errors depends on certain situations. Note that a Type I error (Ŭ) is drawing a conclusion that something is really there i.e. an effect when it actually is not there. Type II error (ɓ) is missing something that is really there. Therefore the seriousness of the error depends on the seriousness of the commission or omission in a particular circumstance e.g. if we are looking for cure for cancer, a Type II error would be very serious because we would be missing a lifesaving treatment. However, if we are considering an expensive drug for the treatment of common cold we will definitely avoid committing Type I error i.e. making false claims over common cold treatment that is self-limiting for an expensive treatment.

Table 16.Relationship between Type 1 and Type 11 errors

		Decision from Sample	
		Accept H_0	**Reject H_0**
True Statement	H_0 True	Correct	Wrong (type-1 error)
	H_0 False	Wrong (Type-11 error)	Correct

3.11 Power of a Test

The power of a test is 1 - the probability of committing a type II error which is given as 1-ϐ. The power of a test is inversely related to the risk of failing to reject a false hypothesis. The greater the ability of a test to eliminate a false hypothesis the greater is its relative power.

4.0 CONCLUSION

In this unit we have explained that tests of significance are mathematical methods used to ascertain if observed difference between sample and population was due to chance. It is usually carried out in tests of proportion or means between sample and population from where the sample was drawn or between two independent samples drawn from different populations or same population groups. The test of significance leads to conclusion that will be inferred on the target population from where the studied sample was drawn. We were able to define test of significance and mentioned various tests statistics used in test of significance. The criteria and process of choosing an appropriate significance test which depend on the type of data collected and the objective of the study were stated. The need for you to consider whether your variables are Nominal (small or large sample), Ordinal or Numerical in order to choose an appropriate significance test was emphasized. We also defined critical region, critical level and critical value as used in test of significance including Type 1 and Type 11 errors

5.0 SUMMARY

In this unit we have learnt:

- that tests of significance are mathematical methods used to verify if observed difference between sample and population occurred by chance.
- that test of significance is usually carried out in tests of proportion or means between samples and population.

- that test of significance leads to drawing conclusion that will be inferred on the target population from where the studied sample was drawn.
- that there are certain criteria and processes required for choosing an appropriate tests.
- important considerations in choosing tests depend on the type of data collected and the objective of the study.
- that there is the need for the consideration of whether variables are Nominal (small or large sample), Ordinal or Numerical in order to choose an appropriate significance test.
- the definitions of critical region, critical level and critical value as used in test of significance.
- the definition of Type 1 and Type 11 errors and their differences and application

6.0 TUTOR-MARKED ASSIGNMENT

1. Define and explain tests of significance.
2. What are the criteria and requirements for test of significance?
3. List the various tests used in test of significance.
4. Define and explain critical region, critical level and critical value as used in test of significance.
5. Define and explain Type 1 and Type 11 errors.
6. When do we commit Type 1 or Type 11 error?

7.0 REFERENCES/FURTHER READING

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York: 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Dorling Kindersley Pvt. Ltd. Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications Ltd. Malden: 3-15.

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

MODULE 9 NON-PARAMETRIC TESTS

Unit 1 Chi-Square test
Unit 2 Other non-parametric tests

UNIT 1 CHI-SQUARE TEST

CONTENTS

1.0 INTRODUCTION

Chi-Square test (X^2) is a Non-parametric test. It does not require the assumption of a normal distribution. Therefore like all other Non-Parametric tests they are tests that do not make assumptions regarding the form of the population from where the samples are drawn because they do not depend on the form of its basic frequency function. They are used when an interval scale cannot be used but the ordering of scores is justified and the sample is small. Therefore, they do not assume any particular functional form for a population distribution but are distribution free methods based on ordered statistical theory. Non-parametric Test hypothesis focuses on nature of distribution alone rather than values. Non-parametric Test is weaker than parametric tests in situations where data satisfies the assumptions. Therefore, they are very simple and easy to use and are more useful for classificatory variables on nominal or ordinal scales of measurement.

2.0 OBJECTIVES

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

- define Chi-Square Test and state the assumptions for its application

- specify the situations and conditions for the use of Chi-Square test
- mention the type of variables applicable in Chi-Square test
- use Chi-Square test in data analysis.

3.0 MAIN CONTENT

3.1 Chi-Square test

Chi-Square test (X^2) is a test of proportion. The Chi-square test is a test statistic that indicates the strength or degree of relationship between two variables. It is a general test which can be used whenever there is an attempt to evaluate whether or not frequencies which have been empirically obtained differ significantly from those which would be expected under a certain set of theoretical assumptions.

Chi-square test (X^2) deals with frequency of categories of one or more variables. It basically determines whether the observed frequencies of given characteristics differ significantly with the expected frequencies under the hypothesis.

3.1.1 Applications of Chi-Square

- **In testing for goodness of fit.** It tests significance of discrepancy between what is theorized and what experiment provides. It ascertains whether the deviation found in the experiment contrasts the theory as a chance difference or theory inadequacy to fit the observed data. So the test evaluates how well the observed frequencies in different categories agree with those expected if the data possessed the same theoretical distribution. It is the determining of how the theory fits into an observation.
- **In testing for the independence of attributes.** The Chi-square test provides the means of testing the hypothesis that the two categorical variables have no relationship existing between them i.e. independent.
- **In the analysis of contingency tables:** Only two-way contingency tables apply in which the rows represent mutually exclusive categories of one variable and the columns also mutually exclusive categories of another variable.
- **In testing proportions.** To find the significant difference in two or more than two proportions. Comparing the values of two Binomial samples even if they are small provided correction factor, Yates correction is applied.
  - Used for a random sample.
  - Used for Qualitative data.

3.1.2 The rules of thumb for Chi-Square test

- When N is greater than 40 use X^2 test with Yateôs correction.
- When N is between 20 and 40 or the expected frequency in each of the four cells is 5 or more use the corrected X^2 test.
- The lowest expected frequency in each cell should not be less than 5 otherwise Yateôs correction is applied thus;

  $\chi^2_c = \Sigma \frac{(|O-E|-0.5)^2}{E}$
- If the smallest expected frequency is less than 5 or if N is less than 20 use Fisherôs exact test.

Note that while the X^2 test approximates the probability, the Fisherôs exact test gives the exact probability of getting a table with values like those obtained or even more extreme.

A measure of the discrepancy existing between the observed and expected frequencies is supplied by the statistic X^2 given by

$$X^2 = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} + \frac{(O_3-E_3)^2}{E_3} + \ldots\ldots\ldots\ldots + \frac{(O_n-E_n)^2}{E_n}$$ for each cell up to n^{th} cells.

óOô is the observed value while óEô is the expected value. The squaring nullifies the negative signs and is standardized by dividing with the expected values.

Degree of freedom (df) is calculated by (Total Row-1) X (Total Column ï 1)

$$\text{Yatesô Correction} = \frac{\chi^2[|O-E|-\frac{1}{2}]^2}{\text{(Total Row-1) X (Total Column ï 1)}}$$

The Chi-square test should not be used if the numbers in the cells are too small. Similarly, for the Chi-square test to be valid, not more than 20% of cells should have expected values of less than 5.

3.1.3 Worked Examples

- In a field trial of a new vaccine each member of a population of 549 persons was randomly assigned into a vaccinated and control groups. Both groups were monitored for a suitable period to see whether they developed the infection or not. The followings are the outcome: Total vaccinated 277, those vaccinated and developed infection 84 and the unvaccinated but developed infection 112. Is there any evidence that the vaccine provides significant protection against the infection?

- It was suspected that cigarette smoking is related to the occurrence of chest disease. A study was carried out on 800 persons out of which 300 were smokers, 68 smokers had chest disease while 88 non-smokers also had chest disease. Is there any association between occurrence of chest disease and smoking of cigarettes?

Solutions
Produce a contingency table for the observed values thus;

Table 17 a. Observed values table

Vaccinated	**Had the Infection**		
	Yes	**No**	**Total**
Yes	84(a)	Not given(b)	277(a+b)
No	112(c)	Not given(d)	Not given(c+d)
Total	**Not given**(a+c)	**Not given**(b+d)	**549**(a+b+c+d)

Table 17b. Observed values table i.e. complete the empty cells for the observed values

Vaccinated	**Had the Infection**		**Total**
	Yes	**No**	
Yes	84	193(277-84)	277
No	112	160(272-112)	272
Total	**196** (84+112)	**353** (193+160)	**549**

Table 17c. Expected values table i.e. calculate and complete the cells for the expected values

Vaccinated	**Had the Infection**		**Total**
	Yes	**No**	
Yes	$\frac{277 \times 196}{549} = 98.9$	$\frac{277 \times 353}{549} = 178.1$	277
No	$\frac{272 \times 196}{549} = 97.1$	$\frac{272 \times 353}{549} = 174.9$	272
Total	**196**	**353**	**549**

$$X^2 = \sum \frac{(O-E)^2}{E} = \frac{(84-98.9)^2}{98.9} + \frac{(193-178.1)^2}{178.1} + \frac{(112-97.1)^2}{97.1} + \frac{(160-174.9)^2}{174.9}$$

$X^2 = 2.24 + 1.25 + 2.29 + 1.27 = 7.07$

*Degree of freedom = (**Total Column-1)(***Total Row-1) = (2-1)(2-1) = 1

Because the calculated value (7.07) is greater than the critical value for 0.05(5%), 0.025(2.5%) and 0.010(1%) significant levels for DF = 1 i.e. 3.84, 5.02, 6.63, respectively, the difference is therefore statistically

significant at these levels i.e. p< 0.05, p<0.025 and p<0.010 or simply stated as p< 0.010. Note that the commonly considered significant levels are 5%, 2.5% and 1%. The conclusion is that the vaccine provides protection against the infection, p<0.010.

Solution
Produce a contingency table for the given observed values thus;

Table 18a Observed values table

	Had chest disease		
Smokers	**Yes**	**No**	**Total**
Yes	68	Not given	300
No	88	Not given	Not given
Total	**Not given**	**Not given**	**800**

Table 18b. Observed values table i.e. complete the empty cells for the observed values

	Had chest disease.		
Smokers	**Yes**	**No**	**Total**
Yes	68	232	300
No	88	412	500
Total	**156**	**644**	**800**

Table 18c. Expected values table i.e. calculate and complete the cells for the expected values

Smokers	**Had chest disease.**		**Total**
	Yes	**No**	
Yes	$\frac{\square\square\square\square\square\square\square\square}{\square\square\square} = 58.5$	$\frac{\square\square\square\square\square\square\square\square}{\square\square\square} = 241.5$	300
No	$\frac{\square\square\square\square\square\square\square\square}{\square\square\square} = 97.5$	$\frac{\square\square\square\square\square\square\square\square}{\square\square\square} = 402.5$	500
Total	**156**	**644**	**800**

$$X^2 = \sum \frac{(\square\square\square)^{\square}}{\square} = \frac{(\square\square\square\square\square\square.\square)^{\square}}{\square\square.\square} + \frac{(\square\square\square\square\square\square\square.\square)^{\square}}{\square\square\square.\square} + \frac{(\square\square\square\square\square\square.\square)^{\square}}{\square\square.\square} + \frac{(\square\square\square\square\square\square\square.\square)^{\square}}{\square\square\square.\square}$$

$X^2 = 1.54 + 0.37 + 0.93 + 0.22 = 3.06$

DF = (TC-1)(TR-1) = (2-1)(2-1) = 1

Because 3.06 is less than the critical value i.e. 3.84 the difference is not statistically significant at 5%, i.e. P > 0.05. Therefore, there is no statistically significant association between smoking and chest disease among the studied population.

Note that in calculating the expected what we are really asking is whether the two categories i.e. control and exposed (vaccinated and not-

vaccinated or smoker and non-smoker) are independent of each other. If they are independent, what frequencies would we expect in each of the cells? Also how different are our observed frequencies from the expected ones?

Therefore if the categories are independent then the probability of a subject being both a control and exposed is P(control) x P(exposed). We then apply the law of joint probability of two independent events.

The expected frequency of an event is equal to the probability of the event times the number of trials = N x P. So the expected number of persons who are both control and exposed is given by;

N x P(control and exposed) = N x P(control) x P(exposed)
From a dummy table below we have

$$= N[\frac{\square\square\square}{\square} \text{ x } \frac{(\square\square\square)}{\square}]$$

$$E = \square\square + \square\square\square \frac{(\square\square\square)}{\square}$$

i.e. for each cell e.g. cell a, multiply the total in the row and column and divide by the grand total.

Table 19. Dummy contingency table

Vaccinated	**Had the Infection**		**Total**
	Yes	**No**	
Yes	A	B	a + b
No	C	D	c + d
Total	**a + c**	**b + d**	**a + b + c + d**

4.0 CONCLUSION

In this unit we learnt that Chi-Square test (X^2) is a test of proportion and a test statistic that indicates the strength or degree of relationship between two variables. It is a general test which can be used whenever there is an attempt to evaluate whether or not frequencies which have been empirically obtained differ significantly from those which would be expected under a certain set of theoretical assumptions. Chi-square test (X^2) deals with frequency of categories of one or more variables. It basically determines whether the observed frequencies of given characteristics differ significantly with the expected frequencies under the hypothesis. He conditions for the application of Chi-Square and its rules of thumb were also discussed.

* Degree of Freedom (DF) is the number of independent values that are free to vary in a data set.

** Total Column (TC) is the total series of values arranged vertically in a contingency table.

*** Total Row (TR) is the total series of values arranged horizontally in a contingency table.

5.0 SUMMARY

In this unit we have learnt:

- the define of Chi-Square test and the assumptions for its application
- the situations and conditions for the use of Chi-Square test
- the type of variables applicable in Chi-Square test
- how to use Chi-Square test in data analysis
- the rules of thumb in Chi-Square application.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define Chi-Square test and state the assumptions for its application.
2. What are the situations and conditions for the use of Chi-Square Test?
3. What are the types of variables applicable in Chi-Square Test?
4. Describe how to use Chi-Square test in data analysis.
5. Highlight the rules of thumb in Chi-Square application.

7.0 REFERENCES/FURTHER READING

Angela, Hebel. (2002). Parametric versus nonparametric statistics when to use them and which is more powerful. Psyphz.psych.wisc.edu/-sha-vs-nonparametricstats.ppt. .

Bartlett, J. E., Kotrlik, J. W and Higgins, C. C. (2001). "Organizational Research: Determining Appropriate Sample Size in Survey Research."*Information Technology, Learning and Performance Journal*. *19* (1) 43- 50.

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York. 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Dorling Kindersley Pvt. Ltd. Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications, Ltd. Boston: 3-15.

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). Lecture Notes on Medical Statistics. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

UNIT 2 OTHER NON-PARAMETRIC TESTS

CONTENTS

1.0 INTRODUCTION

Non-parametric test are test statistics that do not require the assumption of a normal distribution. Non-Parametric tests are tests that do not make assumptions regarding the form of the population from where the samples are drawn because they do not depend on the form of its basic frequency function. They are used when an interval scale cannot be used but the ordering of scores is justified and the sample is small. Therefore, they do not assume any particular functional form for a population distribution but are distribution-free methods based on ordered statistic theory. Non-parametric Test hypothesis focuses on nature of distribution alone rather than values. Non-parametric Test is weaker than parametric tests in situations where data satisfies the assumptions. Therefore, they are very simple and easy to use and are more useful for classificatory variables on nominal or ordinal scales of measurement.

2.0 OBJECTIVES

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

- define Non-parametric tests and state the assumptions for its application
- specify the situations and conditions for the use of non-parametric tests
- discuss the differences between parametric and non-parametric tests
- match different variables with appropriate non-parametric tests
- list the non-parametric tests discussed
- know the usefulness and limitations of non-parametric tests in data analysis

3.0 MAIN CONTENT

3.1 Mann-Whitney U or Wilcoxon test

The Mann-Whitney U or Wilcoxon test requires exactly the same assumptions as the Runs test and follows same procedure. The scores of the two samples are ranked as if from same distribution. Taking each score in the second sample, the number of scores in the first sample which has larger ranks is counted.

The same process is carried out for the scores of the second sample and then the results are added to provide the statistic μ. The sampling distribution of μ can be obtained exactly in a small sample or it can be approximated by a normal curve in the case of larger samples. If μ is either unusually small or unusually large the assumption that the two samples have been drawn from the same population is rejected.

3.2 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is another two sample non-parametric test based on the same assumptions as the previous tests. Unlike them it can be used in instances where there are a large number of ties in the rankings. A number of ties may arise in research when the variables which are ordinal scales may be grouped into a few large categories e.g. in grouping occupations as an ordered variable.

3.3 Kruskal-Wallis test

The Kruskal-Wallis test is a non-parametric test for deciding whether or not two or more samples come from the same population or not. It is an

extension of the Mann Whitney U test. It is used when (distribution) conditions for one-Way ANOVA are not met.

3.4 Fisher's exact test

Fisherôs exact test was developed by R. A. Fisher. It is a statistical test used to determine if there are non-random associations between two categorical variables and used under the same conditions as the Chi-square where the number of cases is small to provide exact rather than approximate probabilities. It is a statistical significance test used in the analysis of contingency tables. Although in practice it is employed when sample size is small, it is valid for all sample sizes. It is one of a class of **exact sample tests**, so called because the significance of the deviation from Null hypothesis i.e. P-value, can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size increases to infinity as with many statistical tests. For hand calculations, the test is only feasible in the case of a 2X2 contingency table.

3.5 Sign test

Sign test is a non-parametric test that uses the sequence of statistic signs of the difference in paired samples for the analysis and is applied in the following situations.

- Used when different pairs are observed under different conditions.
- Used for any pair, each of the two observations matched with respect to relevant extraneous factors.
- Used when any given pair of two observations are being compared.
- It is based on the sign of the deviation

3.6 Mean test

Mean test is a non-parametric test that uses a statistical procedure for testing if the two independent ordered samples differ in their measures of central tendencies.

3.7 Wilcoxon Signed Rank test

The Wilcoxon signed rank test is the non-parametric equivalent to the one-sample Z or t test and the matched pairs test. It is used when we want to make inferences about the mean of one population or the mean difference between two populations in a matched pairs setting.

3.8 Friedman test

This is a non-parametric test that is appropriate whenever the data are measured on, at least, an ordinal scale and can be meaningfully arranged in a two-way classification in a randomized manner.

3.9 Exact Sample test

The entire sampling theory is based on the application of normal tests. However, when the sample size ónô is small, the normal test cannot be applied but exact sample tests can. The exact sample tests can, however, be applied to large samples also, though, the converse is not true. Examples of these exact sample tests are t-test and F-test.

3.10 Runs test

Runs test applies on the assumption that the level of measurement is at least an ordinal scale and the two samples have been drawn from the same continuous population or two identical populations. In the application of Runs test, the data generated from two samples are taken and the scores are ranked from high to low ignoring that they are from different sets of samples. If the Null hypothesis is correct, then the two samples will be mixed so that there is not a long run of cases from the first sample followed by a run of cases from the second. If there are two samples A and B we produce a rank sets such as;

AABBAAABBBABBAABBABA

and not

AAAAAAAAABABBBBBBBBB

In order to test how well the two samples are mixed when ranked, the number of runs are sampled. In the first set there is a run of two As, two Bs, three As etc., with a total runs of 11 while the second set has only four runs. When the number of runs is large as in the first set, the two samples will be mixed so that it is not possible to reject Null hypothesis. Therefore the sampling distribution of runs can be used to establish the critical region used in rejecting the Null hypothesis.

3.11 Test of Significance for difference in proportion

3.11.1 For Single Proportion

Out of 30 children that had measles, 28 survived. If the survival rate for this disease is 85%, test to find out if the infection is more than the known survival rate at 5% significance level.

State H_0: $P_2 = P_1 = 0.85$ i.e. the proportion P_2 of survival observed is equal to the survival rate P_1.

State H_A : $P_2 > 0.85$ i.e. the proportion P_2 of survival observed is more than the survival rate 0.85.

Total survival x = 28, Total infected n = 30, Survival proportion $P_2 = \frac{x}{n} = \frac{28}{30} = 0.93$

Survival rate for the disease was given as 0.85. Therefore $Q_1 = 1 - P_1 = 1 - 0.85 = 0.15$

The test statistic $Z = \frac{(0.93 - 0.85)}{\sqrt{\frac{0.85 \times 0.15}{30}}} = 1.25$

Z at 5% i.e. 0.05 on the table (critical value) is 1.64 for one tail test. Since calculated value 1.25 < 1.64 we therefore do not reject H_0 at this level of significance that the difference is not statistically significant, P > 0.05. In order words the survival rate as observed is not more than the known for the disease in spite of the apparent difference.

3.11.2 For Single Mean

Nine persons have mean haemoglobin of 12.7 mg%. Is the sample drawn from a population with mean 13.6 ± 2.7 mg% at 5% significance level.

State $H_0 : \bar{x} = \mu$ i.e. sample mean = population mean
State $H_A : \bar{x} \neq \mu$ Two tail test

$$Z^* = \left| \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} \right| = \frac{(12.7 - 13.6)}{\frac{2.7}{\sqrt{9}}} = \frac{0.9}{0.9} = 1$$

Z value at 5% is 1.64 which is more than the calculated value i.e. 1< 1.64 therefore we do not reject H_0 i.e. the sample of 9 people was drawn from the population with mean 13.6 ± 2.7 mg%, p> 0.05.

3.11.3 For Two Population Means

The mean blood pressure of 700 and 360 adult males in two communities are 127.56± 10.37 mmHg and 140.78 ± 13.77 mmHg respectively. Is the blood pressure of the males in B community significantly higher than that of community B at 5% significance level.

State $H_0 : \mu_1 = \mu_2$ i.e. No statistically significant difference in their mean blood pressures.

State $H_A : \mu_1 < \mu_2$ i.e. The mean blood pressure 140.78 ± 13.77mmHg is significantly higher than 127.56 ± 10.37 mmHg. This is one-tail test.

$$Z = \frac{\square\square\square\square\square\square}{\sqrt{\frac{\square\square^{\square}}{\square\square} + \frac{\square\square^{\square}}{\square\square}}} = \frac{\square\square\square.\square\square\square\square\square\square\square.\square\square}{\sqrt{\frac{\square\square.\square\square^{\square}}{\square\square\square} + \frac{\square\square.\square\square^{\square}}{\square\square\square}}} = 16.12$$

The calculated value is greater than the tabulated value at 5% significance level 16.12 > 1.64. We therefore reject H_0 because this value falls within rejection region. The difference is statistically significant, $P < 0.05$.

Note that for Test of Association the followings are considered;

- Data analysis involves two categorical variables.
- Values of variables merely define the categories of interest.
- Allocate individuals to corresponding categories.
- Categories are mutually exclusive.
- Size of contingency tables dictated by number of categories.
- Entries in cells of tables are only frequencies (not percentages).

3.12 Advantages of Non-parametric tests

Probability statements obtained from most nonparametric statistics are exact probabilities regardless of the shape of the population distribution from which the random sample was drawn.

- When the sample size is very small there is no alternative to using a nonparametric test.
- Can treat samples made up of observations from several different populations.
- Can treat data which are inherently in ranks as well as data whose seemingly numeral scores have its strength in ranks.
- They are available to treat data which are classificatory.
- Easier to learn and apply compared to parametric tests.

3.13 Shortcomings of Non-parametric tests

- Low Precision and Power of test (the statistical power in probability of rejecting Null hypothesis when it is in fact false and should be rejected).
- Lack of software.
- Wasteful of data.
- Test distributions only.
- Higher-ordered interactions not dealt with.

4.0 CONCLUSION

In this unit we have learnt that Non-parametric tests are tests that do not require the assumption of a normal distribution. The assumptions regarding the form of the population from where the samples are drawn are not also taken into consideration because they do not depend on the form of its basic frequency function. However, it becomes useful when an interval scale cannot be used but the ordering of scores is justified and the sample is small. We also explained that Non-parametric Tests do not assume any particular functional form for a population distribution but are distribution free methods based on ordered statistical theory. That the hypothesis on which non-parametric tests focuses is the nature of distribution alone rather than values. Though they are very simple and easy to use the software is not readily available. The usefulness of non-parametric tests is more for classificatory variables on nominal or ordinal scales of measurement.

5.0 SUMMARY

In this unit we have learnt:

- that Non-parametric tests are tests that do not require the assumption of a normal distribution.
- that non-parametric tests do not make assumptions regarding the form of the population from where the samples are drawn
- that non-parametric tests do not depend on the form of its basic frequency function.
- that non-parametric tests are used when an interval scale cannot be used but the ordering of scores is justified and the sample is small.
- that non-parametric tests are distribution free methods based on ordered statistic theory.
- that non-parametric test hypothesis focuses on nature of distribution alone rather than values.

- that Non-parametric test is weaker than parametric tests in situations where data satisfies the assumptions.
- that non-parametric tests are more useful for classificatory variables on nominal or ordinal scales of measurement.

6.0 TUTOR-MARKED ASSIGNMENT

1. Define Non-parametric tests and state the assumptions for its application.
2. Specify the situations and conditions for the use of non-parametric tests.
3. Discuss the differences between parametric and non-parametric tests.
4. Match different variables with appropriate non-parametric tests.
5. List the non-parametric tests discussed.
6. What are the usefulness and limitations of non-parametric tests in data analysis?

7.0 REFERENCES/FURTHER READING

Angela, Hebel. (2002). Parametric versus nonparametric statistics when to use them and which is more powerful. Psyphz.psych.wisc.edu/-sha-vs-nonparametricstats.ppt.

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York: 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Dorling Kindersley Pvt. Ltd. Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications, Ltd. Malden: 3-15.

Murray, R. S and Larry, J. S. (2006). *Statistics. Theory and Problems of Statistics*. 3rd edition. Tata McGraw-Hill. New Delhi: 1-7.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Habason Nig. Limited. Kano: 32-34.

Syvia, Wassertheil-Smoller. (1990). *Biostatistics and Epidemiology. A primer for health professionals*. Springer-Verlag. New York: 119.

MODULE 10 PARAMETRIC TESTS

UNIT 1 T-TEST

CONTENTS

1.0 INTRODUCTION

Parametric tests are significant tests that are concerned with the population parameters. Sample statistics are obtained in a parametric test to estimate the population parameter. Assumptions in parametric tests include;

- The observations are independent
- Sample data are drawn from a normal distribution.
- Data scores in different groups have homogeneous variance.

Parametric tests make assumptions from the data about the parameters. T-test is a Parametric test and also known as student t- test based on the t-distribution which compares mean values between two groups.

2.0 OBJECTIVES

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

- define parametric tests and their applications
- state the assumptions considered in parametric tests
- list the parametric tests
- match variables and data with appropriate parametric tests

- discuss the parametric calculations and use parametric tests to analyse data
- state the limitations of parametric tests
- define t-test and know the applications
- explain how to use t-test to analyse data.

3.0 MAIN CONTENT

3.1 T-test

This is student t-distribution which compares mean values between two groups. The Student T-test assumes a normal distribution and it is used to estimate averages when only comparatively small samples are involved. Under the circumstance of large sample size it approximates to Z. T-test was introduced by W.S. Gossett writing under the pseudonym 'Student', then as the firm he worked for did not permit their staff publishing articles in their own names.

3.1.1 Applications

- To test if the sample mean ( ) differs significantly from population mean (μ).
- To test the significance difference between two sample means.
- To test the significance of sample correlation coefficient.
- Widely used in medical research.

3.1.2 Assumptions

- That the two groups compared have similar variances.
- That the population from which the sample is drawn is normal.
- That the sample is randomly distributed.
- That each score is measured separately in each group and are independently drawn.
- That the parent population's standard deviation (ů) is unknown.
- Data is measured on at least an interval scale.

3.1.3 Criteria for applying t-test

- In Quantitative data.
- In Random samples.
- In sample size less than 30
- In variables normally distributed.

For Single Mean (One sample t-test)

Worked Examples

A random sample of 10 students has the following Intelligence Quotient (IQ) 67,110,115,75,63,117,120,115,100 and 97. Is this sample drawn from a population of Science students with IQ of 100.

State H_0: $\bar{x} = \mu$ i.e. $\bar{x} = 100$ (the sample IQ is the same with the Science studentôs IQ).

State H_A : $\bar{x} \neq \mu$ i.e. $\bar{x} \neq 100$ (Not the same, a two tail test)

We can calculate $\bar{x}$ and SD from the above data thus;

$\bar{x} = \sum \frac{x}{n} = \frac{976}{10} = 97.6$

$SD = \frac{\sqrt{(x - \bar{x})^2}}{n-1} = 21.85$

$t = \frac{\bar{x} - \mu}{\frac{SD}{\sqrt{n}}} = \frac{97.6 - 100}{\frac{21.85}{\sqrt{10}}} = 0.34$

Degree of Freedom (df) = n -1 = 10 -1 = 9 The critical value at 5% significance level and df of 9 is 2.62 on the t-test distribution table. Calculated value 0.34 < 2.62 hence we do not reject H_0 that the sample is drawn from the population of medical students with IQ of 100, p> 0.05.

- **For comparison of two mean values (independent sample t-test)**

Independent samples: $t = \frac{(\bar{x}_1 - \bar{x}_2)}{SE(\bar{x}_1 - \bar{x}_2)}$

$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ when the variances are not equal

$SE(\bar{x}_1 - \bar{x}_2) = s^2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ when the variances are equal

where the pooled variance $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + (\bar{x}_1 - \bar{x}_2)^2}{n_1 + n_2 - 2}$

- **Worked example for two independent population means**

Independent sample of age onset of symptoms in lung cancer of 12 female and 13 male patients is given below. Is there any significant difference in their mean age of onset of symptoms?

Table 20. Independent populations of males and females

Sex	Age (yrs)
Female (F)	58, 52,50,49,56,52,54,48,41,37,67,70
Male (M)	26,41,57,66,36,55,41,61,53,50,52,37,50

State H_0: $\bar{x}_{F}$ $\bar{x}_{M}$ No difference

State H_A : $\bar{x}_{F}$ $\bar{x}_{M}$ There is a difference

Calculate the statistics from the above table as follows;

$\bar{x}_1 = 12$ $\bar{x}_2 = 13$

$s_1^2 = 52.83$ $s_2^2 = 48.08$

$n_1 = 93$ $n_2 = 120$

Substitute for the values in t-test formula (for the unequal variance) below;

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 1.15$$

From the table of standard normal distribution 1.16> 1.15 calculated value so we do not reject H_0that there is no difference in their mean age at the onset of symptoms, P > 0.05.

Suppose in a community the mean heights of 60 women with normal deliveries and 52 with Caesarean sections are 156± 3.1cm and 154±2.8 cm respectively. Test the Null hypothesis that there is no association with height of the women and delivery outcome.

Calculate the difference in their mean values i.e. 156 ï 154 = 2 cm
Substitute the values thus;

$$t = \frac{2}{\sqrt{\frac{3.1^2}{60} + \frac{2.8^2}{52}}} = 3.6$$

df = 60 + 52 ï 2 = 110

t-value at p=0.05 for df = 110(120 on the table) is 1.98

Since the calculated value is higher than the table value we reject Null hypothesis and say that there is an association.

4.0 CONCLUSION

In this unit we have learnt that Parametric Tests are significant tests that are concerned with the population parameters. That sample statistics are obtained in a parametric test to estimate the population parameter. T-Test is one of the commonly used parametric tests with the following assumptions; that the observations are independent, sample data are drawn from a normal distribution and that Data are scores in different groups have homogeneous variance. Parametric tests also make assumptions from the data about the parameters. T-test is also known as student t-distribution which compares mean values between only two groups.

5.0 SUMMARY

In this unit we have been able to:

- define parametric tests and the applications
- state the assumptions considered in parametric tests
- list the parametric tests
- match variables and data with appropriate parametric tests
- explain the parametric calculations and how to use parametric tests to analyze data
- mention the limitations of parametric tests
- define T-Test and mention the assumptions and applications in its use
- describe how to use T-test to analyse data with worked examples

6.0 TUTOR-MARKED ASSIGNMENT

1. Define parametric tests and mention the applications.
2. State the assumptions considered in parametric tests.
3. List the parametric tests.
4. Match variables and data with appropriate parametric tests.
5. Explain the parametric calculations and how to use parametric tests to analyze data.
6. Mention the limitations of parametric tests.
7. What is t-test?
8. Mention the assumptions and applications of t-test?
9. Describe how to use t-test to analyse data with worked examples.

7.0 REFERENCES/FURTHER READING

Angela, Hebel. (2002). Parametric versus nonparametric statistics when to use them and which is more powerful. Psyphz.psych.wisc.edu/-sha-vs-nonparametricstats.ppt. .

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York: 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Dorling Kindersley Pvt. Ltd. Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications, Ltd. Malden: 3-15.

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

UNIT 2 ANALYSIS OF VARIANCE (ANOVA) AND COVARIANCE

CONTENTS

1.0 INTRODUCTION

Analysis of Variance (ANOVA) is a parametric test of statistical technique whereby the total variation present in a set of data is partitioned into two or more components. Associated with each of these components is a specific source of variation, so that in the analysis it is possible to ascertain the magnitude of the contributions of each of these sources to the total variation. ANOVA is the appropriate technique for analyzing continuous variables when there are three or more groups to be compared e.g. comparing the blood pressure reduction effects of three drug therapeutic trial.

ANCOVA is the method of test statistics used when some of the independent variables are categorical and others continuous. It is used to achieve statistical control of error when experimental control is not feasible. ANCOVA will then adjust the analysis in two ways;

- Reduce the estimates of experimental error.
- Adjusts treatment effects with respect to the covariate.

In most cases the scores on the covariate are collected before the experimental treatment e.g. pretest scores or exam scores etc. In some

experiments the scores on covariate are collected after the experimental treatment e.g. anxiety, motivation, depression etc. In ANCOVA variance is partitioned into three basic components viz. Effect, Error, Covariate.

2.0 OBJECTIVES

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

- define and explain the application of ANOVA and ANCOVA.
- state the assumptions considered in ANOVA and ANCOVA
- state what types of observations are used in ANOVA and ANCOVA
- explain what kind of population distribution we draw data for parametric tests?
- explain what kind of different groups variance is considered in parametric test?

3.0 MAIN CONTENT

3.1 Analysis of Variance (ANOVA)

The principles involved in the ANOVA are the same as those in the t-test. Essentially, it is to know if the variability of all the groupsô mean is substantially greater than the variability within each of the groups around their own means. We calculate a quantity known as the between-groups variance, which is the variability of the group means around the mean of all the data. We calculate another quantity called the within-groups variance which is the variability of the scores within each group around its own mean. One of the assumptions of the Analysis of Variance is that the extent of the variability of individuals within groups is the same for each of the groups, so we can pool the estimate of the individuals within group variances to obtain a more reliable estimate of overall within groups variance. If there is as much variability of individuals within the groups as there is variability of means between the groups they probably come from same population which would be consistent with the hypothesis of no true difference among means i.e. we could not reject the Null hypothesis of no difference among means. The ratio of the between-groups variance to the within-groups variance is known as the F-ratio.

The reason for doing an ANOVA is to see if there is any difference between groups on some variables. Underlying the valid use of Analysis of Variance as a tool of statistical inference are set of fundamental assumptions.

For example, you might have data on student performance in non-assessed tutorial exercises as well as their final grading. You are interested in seeing if tutorial performance is related to final grade. ANOVA allows you to break up the group according to the grade and then see if performance is different across these grades. ANOVA is available for both parametric (score data) and non-parametric (ranking/ordering) data.

3.1.1 One-way (ANOVA) measurement between groups

The example given above is called a one-way between groups model. You are looking at the differences between the groups. There is only one grouping (final grade) which you are using to define the groups. This is the simplest version of ANOVA. This type of ANOVA can also be used to compare variables between different groups - tutorial performance from different intakes.

3.1.2 One-way (ANOVA) repeated measurements

A one way repeated measures ANOVA is used when you have a single group on which you have measured something a few times. For example, you may have a test of understanding the classes you took e.g. you give this test at the beginning of each lesson and at the end of the lesson and then at the end of the subject. You would use one-way repeated measures ANOVA to see if studentôs performance on the tests changed over time.

3.1.3 Two-way (ANOVA) measurement between groups

A two-way between groups ANOVA is used to look at complex groupings. For example, the grades by tutorial analysis could be extended to see if overseas students performed differently to local students. What you would have from this form of ANOVA is;

- The effect of final grade,
- The effect of overseas versus local,
- The interaction between final grade and overseas/local,

Each of the main effects is one-way test. The interaction effect is simply asking "is there any significant difference in performance when you take final grade and overseas/local acting together".

3.1.4 Two-way (ANOVA) repeated measurements

This version of ANOVA simply uses the repeated measures structure and includes an interaction effect. In the example given for one-way

between groups, you could add gender and see if there was any joint effect of gender and time of testing i.e. do males and females differ in the amount they remember/absorb over time.

ANOVA is available for score or interval data as parametric ANOVA. This is the type of ANOVA you do from the standard menu options in a statistical package. The non-parametric version is usually found under the heading "Nonparametric test". It is used when you have ranked or ordered data.

You cannot use parametric ANOVA when your data is below interval measurement. Where you have ***categorical*** data you do not have an ANOVA method - you would have to use Chi-square which is about interaction rather than about differences between groups.

3.1.5 Procedure

What ANOVA looks at is the way groups differ internally versus what the difference is between them. To take the above example:

- ANOVA calculates the mean for each of the final grading groups on the tutorial exercise figure i.e. the group means.
- It calculates the mean for all the groups combined i.e. the overall mean.
- Then it calculates, within each group, the total deviation of each individual's score from the group mean i.e. within group variation.
- Next, it calculates the deviation of each group mean from the overall mean i.e. between group variation.
- Finally, ANOVA produces the F statistic which is the ratio between group variation to the within group variation.

If the between group variation is significantly greater than the within group variation, then it is likely that there is a statistically significant difference between the groups. The statistical package will tell you if the F ratio is significant or not.

All versions of ANOVA follow these basic principles but the sources of variation get more complex as the number of groups and the interaction effects increase. The calculations required by analysis of variance are lengthier and more complicated than the other significance tests. For this reason the computer assumes an important role in analysis of variance using appropriate computer statistical packages.

3.2 Analysis of Covariance (ANCOVA)

ANCOVA is the method of test statistics used when some of the independent variables are categorical and others continuous. It is used to achieve statistical control of error when experimental control is not feasible. ANCOVA will then adjust the analysis in two ways;

- Reduce the estimates of experimental error.
- Adjust treatment effects with respect to the covariate.

In most cases the scores on the covariate are collected before the experimental treatment e.g. pretest scores or exam scores etc. In some experiments the scores on covariate are collected after the experimental treatment e.g. anxiety, motivation, depression etc. In ANCOVA variance is partitioned into three basic components viz. Effect, Error, Covariate.

3.2.1 Assumptions in the use of ANCOVA

- All the assumptions that apply to between groups in ANOVA.
- The assumption of linear regression.
- The assumption of homogeneity of regression coefficients.

3.3 Comparing mean value for some variables of interest

For two samples;
Parametric test use -t-test for independent samples.
Nonparametric test use -Wald-Wolforvitz Runs test
-Kolmogorov-Smirnov two sample test.
- Mann-Whitney U-test

Difference between independent groups
For multiple groups;
Parametric test use -ANOVA
-MANOVA
Nonparametric test use -Kruskal-Wallis analysis of ranks
-Median test

3.4 Difference between dependent groups

For comparing two variables measured in the same sample.
Parametric test use; -t-test for dependent samples.
Nonparametric test use; -Sign test.
-Wilcoxonôs matched pairs test.

For more than two variables measured in same sample.
Parametric test use -Repeated measures ANOVA
Nonparametric test -Friedmanôs two way ANOVA.
-Cochran Q

3.5 Relationships between variables

For two variables of interest that are categorical.
Parametric test use -Correlation coefficient (r)
Nonparametric test -Spearman R
-Kendal Tau
-Coefficient Gamma
-Chi Square
-Phi Coefficient
-Fisher's exact test
-Kendall Coefficient of concordance.

Table 21. Parametric and Nonparametric data and test statistics.

Scale	Process	Data	Test Statistics	
Ratio	Equal intervals True zero Ratio relationship	Parametric	Descriptive	Inferential
Interval	Equal intervals No true zero		Mean Standard deviation Pearsonôs *r*	T-test ANOVA
Ordinal	Ranked in order	Nonparametric	Median Quartile deviation Spearmanôs *rho*	Mann-Whitney Wilcoxon
Nominal	Classified and Counted		Mode	Chi-Square Sign

4.0 CONCLUSION

Analysis of Variance (ANOVA) and Analysis of Covariance (ANCOVA) are Parametric tests that significant tests concerned with the population parameters. Sample statistics are obtained in a parametric test to estimate the population parameter. Assumptions in parametric test usually includes, that the observations are independent, that the sample data are drawn from a normal distribution and that the data are scores in different groups have homogeneous variance. Therefore, parametric

tests make assumptions from the data about the parameters. Some of the commonly used parametric tests are T-test, Analysis of Variance (ANOVA) and Analysis of Co-Variance (ANCOVA).

5.0 SUMMARY

In this unit we have learnt:

- ANOVA and ANCOVA are parametric tests that are also significant tests that are concerned with the population parameters.
- that sample statistics are obtained in a parametric test such as ANOVA and ANCOVA to estimate the population parameter.
- that parametric test observations as in ANOVA and ANCOVA are independent
- that in parametric tests such as ANOVA and ANCOVA the sample data are drawn from a normal distribution
- that in use of parametric tests such as ANOVA and ANCOVA the data are scores in different groups that have homogeneous variance.
- that parametric tests such as ANOVA and ANCOVA make assumptions from the data about the parameters.
- that the commonly used parametric tests are T-test, Analysis of Variance (ANOVA) and Analysis of Co-Variance (ANCOVA).

6.0 TUTOR-MARKED ASSIGNMENT

1. Define and explain the application of ANOVA and ANCOVA.
2. What are the assumptions considered in ANOVA and ANCOVA?
3. Explain what type of observations are used ANOVA and ANCOVA.
4. From what kind of population distribution do we draw data for ANOVA and ANCOVA?
5. What kind of different group variance is considered in ANOVA and ANCOVA?
6. Describe how ANOVA and ANCOVA test statistics are used in data analysis.

7.0 REFERENCES/FURTHER READING

Angela, Hebel. (2002). Parametric versus nonparametric statistics when to use them and which is more powerful. Psyphz.psych.wisc.edu/-sha-vs-nonparametricstats.ppt. .

Charles, S. D. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag. New York: 15.

Hannagan, T. (1997). *Mastering Statistics*. 3rd edition. Palgrave Master Series. New York: 1-8.

Jerrold, H. Z. (2006). *Biostatistical Analysis*. 4th edition. Dorling Kindersley Pvt. Ltd. Indore: 32-39.

Kirkwood, B. R and Sterne, J. A. C. (2010). *Essential Medical Statistics*. 2nd edition. Blackwell Scientific Publications Ltd. Malden: 3-15.

Mahajan, B. K. (2010). *Methods in Biostatistics for Medical Students and Research Workers*.7th edition. Jaypee Brothers Medical Publishers Ltd. New Delhi: 33-52.

Ogbonna, C. (2016). *The Basics in Biostatistics, Medical Informatics and Research Methodology*. 3 in 1 Book. Revised Ed. Yakson Printing Press. Jos: 3-128.

Petrie, A. (1986). *Lecture Notes on Medical Statistics*. Blackwell Scientific Publications Ltd. Edinburgh: 40-46.

Singha, P. (1996). *An Introductory Text on Biostatistics*. 2nd edition. Habason Nig. Limited. Kano: 32-34.

Appendix I

Table for determining minimum returned sample size for a given population size for continuous and categorical data.

POPULATION SIZE	MINIMUMSAMPLE SIZE					
	Continuous data for margin of error = 0.03			Categorical data for margin of error = 0.05		
	alpha= .10 Z = 1.65	alpha = .05 Z = 1.96	alpha = .01 Z = 2.58	p = .50 Z = 1.65	p= .50 Z = 1.96	p = .50 Z = 2.58
100	46	55	68	74	80	87
200	59	75	102	116	132	154
300	65	85	123	143	169	207
400	69	92	137	162	196	250
500	72	96	147	176	218	286
600	73	100	155	187	235	316
700	75	102	161	196	249	341
800	76	104	166	203	260	363
900	76	105	170	209	270	382
1,000	77	106	173	213	278	399
1,500	79	110	183	230	306	461
2,000	83	112	189	239	323	499
4,000	83	119	198	254	351	570
6,000	83	119	209	259	362	598
8,000	83	119	209	262	367	613
10,000	83	119	209	264	370	623

Table developed by Bartlett, Kotrlik and Higgins

Appendix II

Table of Random Numbers

40 75 30 74 05 72 32 19 95 37 80 21 66 04 15 32 30 28 22 53 88 74 19 56 17
30 99 22 70 87 31 17 08 73 17 13 96 43 74 57 71 01 76 65 72 48 62 15 59 28
02 72 63 12 50 64 75 71 30 87 89 17 23 30 00 16 01 37 11 31 10 25 06 63 22
09 74 65 74 23 98 20 37 98 58 99 53 56 05 34 44 37 06 75 56 28 85 18 36 89
10 19 83 99 45 02 14 55 05 94 24 45 23 37 54 74 30 43 82 39 64 93 32 15 29

74 13 55 59 05 45 32 39 84 51 60 11 36 64 61 39 82 97 42 10 66 11 88 78 46
84 52 14 96 06 41 18 55 83 44 83 23 41 44 28 79 01 62 98 46 76 25 27 76 48
23 72 86 82 80 57 96 27 48 51 34 75 50 54 02 01 08 54 20 73 55 94 22 18 82
81 22 51 33 44 67 75 19 59 29 88 72 81 32 92 77 27 73 92 45 34 61 25 05 16
02 34 64 55 31 30 01 08 22 67 28 64 43 25 78 66 64 84 88 87 10 05 22 16 92

27 16 61 45 62 70 36 69 03 79 50 64 26 76 11 27 76 91 93 45 76 84 07 94 27
80 21 66 04 15 64 75 71 30 87 34 61 25 05 16 02 14 55 05 94 84 52 14 96 06
71 94 78 09 08 00 90 36 87 69 65 84 40 62 81 71 05 60 48 65 22 24 63 02 76
04 22 76 32 80 86 72 93 83 63 96 50 84 02 77 35 63 40 69 52 77 34 40 77 65
23 01 71 61 37 31 96 19 68 07 11 13 63 24 47 01 04 41 35 27 81 97 31 17 89

50 73 06 39 73 45 80 89 69 92 71 82 70 81 33 22 02 66 22 47 68 45 81 50 20
03 03 51 04 64 50 70 60 26 87 57 00 77 75 33 90 51 71 24 69 75 75 65 95 33
75 98 56 87 90 19 43 10 64 11 86 33 76 92 25 12 74 95 04 75 75 06 09 85 74
93 16 38 20 19 18 13 50 71 57 23 46 29 66 39 57 98 65 05 39 24 06 47 20 70
00 43 17 47 14 73 09 73 72 17 83 60 65 09 76 93 55 82 04 93 06 79 83 16 27

64 50 82 70 23 74 20 89 62 74 35 00 85 22 45 1142 81 84 19 67
45 51 41 73
44 24 74 92 82 13 92 55 05 62 91 36 80 87 99 25 78 36 90 50 76
71 09 47 40
26 09 80 81 72 60 58 93 60 62 76 33 77 65 35 22 60 74 14 12 09
43 76 92 79
71 49 96 71 49 63 23 87 66 03 10 74 01 26 40 00 28 52 67 66 07
78 48 74 79
14 94 04 41 98 87 27 41 58 12 01 30 87 67 72 05 62 78 64 78 29
78 72 33 45

37 31 28 75 75 34 61 25 05 16 11 66 32 93 01 74 19 14 31 56 92
62 26 06 69
36 80 53 67 95 09 29 64 74 36 74 49 84 66 16 34 35 09 46 77 21
77 64 97 62
97 74 58 93 15 90 94 03 65 64 42 85 95 60 00 16 32 99 35 14 94
86 83 89 60
22 35 14 05 57 09 26 73 30 34 71 68 44 96 29 05 10 12 45 7292 05
14 30 82
06 90 87 84 44 36 22 56 20 97 60 92 47 83 11 13 33 23 34 24 04
33 27 82 66

86 36 50 45 77 07 36 17 23 29 04 48 96 60 01 66 12 41 35 83 26
49 92 91 86
51 21 84 81 49 10 35 09 87 75 10 45 44 69 88 87 57 98 33 09 48
16 44 72 22
86 09 74 30 59 60 47 08 47 41 62 65 44 79 77 93 16 21 40 76 43
97 44 74 91
96 09 02 52 03 62 91 46 48 76 90 76 31 69 17 66 50 09 39 49 01
08 54 20 73
70 16 32 99 69 60 26 04 43 61 09 74 65 74 23 67 62 23 74 57 88
45 75 23 37

64 93 32 15 29 44 48 76 58 68 42 03 95 62 67 79 01 62 98 46 95
31 96 60 18
04 23 06 30 27 82 25 28 53 13 13 89 23 88 30 63 58 27 96 73 40
01 78 50 05
99 69 41 08 09 37 55 01 61 59 82 32 45 29 17 34 69 12 10 65 48
98 02 70 73
59 17 16 04 51 51 39 51 55 05 68 88 09 80 01 69 12 97 38 36 05
07 18 99 83
57 03 21 00 04 46 78 08 83 74 01 69 74 69 21 80 15 88 25 76 91
84 81 45 03

26 87 94 08 38 21 32 43 02 37 51 67 04 69 88 28 42 29 46 50 69
41 29 86 14

65 12 14 15 85 34 52 03 81 54 30 33 02 87 68 32 86 45 62 13 19 39 82 43 20
14 06 77 39 33 95 50 69 19 43 31 97 79 48 40 04 29 97 46 03 46 46 26 90 30
13 24 91 13 98 50 02 32 24 48 48 74 43 37 80 32 92 85 43 29 19 33 57 28 58
36 33 29 39 22 43 21 04 30 14 23 52 06 74 24 63 43 33 83 04 10 53 89 14 98

Appendix III

Percentage points of the χ^2 distribution.

DF	**0.975**	**0.900**	**0.750**	**0.500**	**0.250**	**0.100**	**0.050**	**0.025**	**0.010**	**0.001**
1	-	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	10.83
2	0.05	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	13.82
3	0.22	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	16.27
4	0.48	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	18.47
5	0.83	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	20.52
6	1.24	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	22.46
7	1.69	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	24.32
8	2.18	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	26.12
9	2.70	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	27.88
10	3.25	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	29.59
11	3.82	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	31.26
12	4.40	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	32.91
13	5.01	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	34.53
14	5.63	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	36.12
15	6.27	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	37.70
16	6.91	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	39.25
17	7.56	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	40.79
18	8.23	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	42.31
19	8.91	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	43.82

20	9.59	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	45.32
21	10.28	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	46.80
22	10.98	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	48.27
23	11.69	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	49.73
24	12.40	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	51.18
25	13.12	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	52.62
26	13.84	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	54.05
27	14.57	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	55.48
28	15.31	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	56.89
29	16.05	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	58.30
30	16.79	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	59.70
40	24.43	29.05	33.66	39.34	45.62	51.80	55.76	59.34	63.69	73.40
50	32.36	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	86.66
60	40.48	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	99.61
70	48.76	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	112.32
80	57.15	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	124.84
90	65.65	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	137.21
100	74.22	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	149.45

Appendix IV
Percentage points of the Student t-distribution

DF	**α**	**0.45**	**0.40**	**0.35**	**0.30**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**
		0.025	**0.010**	**0.005**						
v	**2α**	**0.90**	**0.80**	**0.70**	**0.60**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**
		0.050	**0.020**	**0.010**						
1		0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314
		12.706	31.821		63.657					
2		0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920
4.303		6.965		9.925						
3		0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353
3.182		4.541		5.841						
4		0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132
2.776		3.747		4.604						
5		0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015
2.571		3.365		4.032						
6		0.130	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943
2.447		3.143		3.707						
7		0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895
2.365		2.998		3.499						
8		0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860
2.306		2.896		3.355						
9		0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833
2.262		2.821		3.250						
10		0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812
2.228		2.764		3.169						
11		0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796
2.201		2.718		3.106						
12		0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782
2.179		2.681		3.055						
13		0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771
2.160		2.650		3.012						
14		0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761
2.145		2.624		2.977						
15		0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753
2.131		2.602		2.947						
16		0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746
2.120		2.583		2.921						
17		0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740
2.110		2.567		2.898						
18		0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734
2.101		2.552		2.878						

19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729
2.093	2.539	2.861							
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725
2.086	2.528	2.845							
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721
2.080	2.518	2.831							
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717
2.074	2.508	2.819							
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714
2.069	2.500	2.807							
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711
2.064	2.492	2.797							
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708
2.060	2.485	2.787							
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706
2.056	2.479	2.779							
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703
2.052	2.473	2.771							
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701
2.048	2.467	2.763							
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699
2.045	2.462	2.756							
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697
2.042	2.457	2.750							
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684
2.021	2.423	2.704							
50	0.126	0.255	0.388	0.528	0.679	0.849	1.047	1.299	1.676
2.009	2.403	2.678							
100	0.126	0.254	0.386	0.526	0.677	0.845	1.042	1.290	1.660
1.984	2.364	2.626							
200	0.126	0.254	0.386	0.525	0.676	0.843	1.039	1.286	1.653
1.972	2.345	2.601							
Ð	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645
1.960	2.326	2.576							

Appendix V

The Correlation Coefficient Table Probability. The values of the correlation, *r*, at different levels of significance when testing H_o: ρ = 0. For *two-tailed* test, significance is achieved at the specified level if the absolute value of the sample correlation coefficient based on *n* pairs of observations exceeds the tabulated value where, $v = n - 2$. For *one-tailed* test, the significance level is halved.

DF, *v*	2α	0.100	0.050	0.020	0.010	0.001
1		0.987	0.997	0.999	0.999	0.999
2		0.900	0.950	0.980	0.990	0.999
3		0.805	0.878	0.934	0.959	0.991
4		0.729	0.811	0.882	0.917	0.974
5		0.669	0.755	0.833	0.875	0.951
6		0.621	0.707	0.789	0.834	0.925
7		0.582	0.666	0.750	0.798	0.898
8		0.549	0.632	0.716	0.765	0.872
9		0.521	0.602	0.685	0.735	0.847
10		0.497	0.576	0.658	0.708	0.823
11		0.476	0.553	0.634	0.684	0.801
12		0.457	0.532	0.612	0.661	0.780
13		0.441	0.514	0.592	0.641	0.760
14		0.426	0.497	0.574	0.623	0.742
15		0.412	0.482	0.558	0.606	0.725
16		0.400	0.468	0.543	0.590	0.708
17		0.389	0.456	0.529	0.575	0.693
18		0.378	0.444	0.516	0.561	0.679

19		0.369	0.433	0.503	0.549
	0.665				
20		0.360	0.423	0.492	0.537
	0.652				
25		0.323	0.381	0.445	0.487
	0.597				
30		0.296	0.349	0.409	0.449
	0.554				
35		0.275	0.325	0.381	0.418
	0.519				
40		0.257	0.304	0.358	0.393
	0.490				
45		0.243	0.288	0.338	0.372
	0.465				
50		0.231	0.273	0.322	0.354
	0.443				
60		0.211	0.250	0.295	0.325
	0.408				
70		0.195	0.232	0.274	0.302
	0.380				
80		0.183	0.217	0.257	0.283
	0.357				
90		0.173	0.205	0.242	0.267
	0.338				
100		0.164	0.195	0.230	0.254
	0.321				

Appendix VI

Standard Normal Distribution (One-tailed)

Z	**0.0000**	**0.0100**	**0.0200**	**0.0300**	**0.0400**	**0.0500**	**0.0600**	**0.0700**	**0.0800**	**0.0900**
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4483	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985

1.3	0.0968	0.0951	0.0934	0.1918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0219	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019

2.9
0.00190.00180.00180.00170.00160.00160.00150.00150.00140.0
014
3.0
0.00130.00130.00130.00120.00120.00120.00110.00110.00100.0
010
3.1 0.00090.0009 0.0009 0.0009 0.0008 0.0008 0.0008
0.0008 0.0007 0.0007
3.2 0.0007 0.0007 0.0006 0.0006 0.0006 0.0006 0.0006
0.0005 0.0005 0.0005
3.3 0.0005 0.0005 0.0005 0.0005 0.0004 0.0004 0.0004
0.0004 0.0004 0.0004
3.4 0.0003 0.0003 0.0003 0.0003 0.0003 0.0003 0.0003
0.0003 0.0003 0.0003
3.5 0.0002 0.0002 0.0002 0.0002 0.0002 0.0002 0.0002
0.0002 0.0002 0.0002

The first vertical column shows the Z values in one decimal point e.g. for Z = 2.0 corresponds to 0.0228 etc. The first row values to the right of Z shows the Z values up to two decimal points e.g. for Z = 2.01 corresponds to 0.0222 etc. Also to find the area or percentage of observations lying beyond the Z value of 1.96 first find 1.9 in the Z column at the extreme left and then read across the first row of 0.06. The intercept of the two values reads 0.0250 i.e. 2.5% of the observations lie within this area.

Appendix VII

Percentage points of the Standard Normal Distribution (z)

Percentage Points

P-value	***One-sided***	***Two-sided***
0.5	0.00	0.67
0.4	0.25	0.84
0.3	0.52	1.04
0.2	0.84	1.28
0.1	1.28	1.64
0.05	1.64	1.96
0.02	2.05	2.33
0.01	2.33	2.58
0.005	2.58	2.81
0.002	2.88	3.09
0.001	3.09	3.29
0.0001	3.72	3.89