COURSE GUIDE

STT102
INTRODUCTORY STATISTICS

Course Team Dr. Paye V. Folarin (Course Writer) -University of Ibadan, Ibadan
Dr. Akeem B. Disu (Course Editor) - NOUN
Oyewole A. Oyelami (Course Reviewer) - 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

Printed

ISBN: 978-058-643-1

CONTENTS **PAGE**

INTRODUCTION

STT102 - Introductory Statistics is a second semester, *two credit*, basic course designed for students with no previous knowledge of statistics. It is intended to initiate students to the subject at an introductory stage so that they acquire fundamental concept required in gaining deeper understanding of social, health and related issues.

The course consists of *twenty* units divided into *four modules* of content of *five units* each. It involves basic principles in collection, compilation, analysis, presentation of data and the drawing of conclusions from statistical analysis. The material has drawn on several practical examples from the local environment so that the relevance of theory to the practical situation may be appreciated. Extensive use of the media and several case studies in the social and health sectors from information collection agencies and government data support the student's study.
Since computers are being used to an increasing extent in the hospital services, and in view of the fact that they, are ideal for performing statistical calculations and analysis, the student is advised to be computer literate. However, the need to write one's computer programs does not arise since they already exist for almost all aspects of statistics. The student requires no pre-requisites for the course.

COURSE COMPETENCIES

The general objective of Introductory Statistics is to introduce the student to the basic principles and applications of statistics in nursing practice and research. During the course, the students will learn several statistical concepts and techniques and how to employ them in making generalizations and decisions on social, health and related issues.
In this course, the student will acquire many well-organized statistical procedures that will endow him with strategies needed in solving positively many medical problems.

The aim of the course is to introduce the student to the statistical process and methods in common use.
It will be achieved by:

- Introducing you to fundamental statistical concepts and procedures. Illustrating how these principles can be applied to issues in nursing practice and research.

- Explaining to you how data and case studies from information-collection agencies and government sources may be employ to solve current health problems.

- Giving you the grounding in the ability to generalize and decide positively on situations arising in the health sector.

COURSE OBJECTIVES

At the end of this course, you should be able to attend the following:

- Collect adequate and reliable statistical information.
- Present the data in forms in which the main characteristics are easily Understood.
- Organize and summarize the information collected.
- Treat the data scientifically i.e., analyze the main features of the data.
- Deduce meaningful conclusions or relationships from the statistical analysis.
- Measure the reliability of the conclusions about a population based on information of the population.

WORKING THROUGH THIS COURSE

You are required to read the study units, set books and other materials provided by the National Open University to complete the course. You will also need to work through practical and self- assessed exercises and submit assignments for assessment purposes. The course will take you about 60 hours to complete at the end of which you will write a final examination.

This course consists of the following *twenty* study units:

Module 1

Unit 1 Aims of the Statistical Method
Unit 2 Collection of Data: Sampling
Unit 3 Collection of Data: Bias
Unit 4 Collection of Data: Forms of Record
Unit 5 Presentation of Data

Module 2

Unit 1 Frequency Distribution
Unit 2 Cumulative Distribution
Unit 3 Measures of Location
Unit 4 Measures of Dispersion
Unit 5 Correlation

Module 3

Unit 1	Regression
Unit 2	Simple Concepts of Probability
Unit 3	Relationship between Population and Sample
Unit 4	Normal Distribution
Unit 5	Sampling Distribution of the Mean and the Central Limit Theorem

Module 4

Unit 1	Mean Estimation
Unit 2	Fundamentals of Hypothesis Test
Unit 3	Hypothesis Test for one Population Mean when Standard Deviation is Known
Unit 4	Classical Approach vs P-Value Approach to Hypothesis Testing
Unit 5	Measures of Morbidity

The first fourteen units deal with fundamental principles of statistics while the last six units involve application of these principles. Each study unit will take about four weeks work. It involves specific objectives, how to study the reading materials, references to set books and other related sources and summaries of vital points and ideas. The unit direct you to work on exercises related to the require reading and to carry out practical computer exercises where appropriate. A number of self-tests are associated with each unit. These tests give you an indication of your progress. The exercises as well as the tutor-marked assignments will help you in achieving the stated learning objectives of each unit and of the course.

PRESENTATION SCHEDULE

The weekly activities are presented in Table 1 while the required hours of study and the activities are presented in Table 2. This will guide your study time. You may spend more time in completing each module or unit.

Table I: Weekly Activities

Week	**Activity**
1	Orientation and course guide
2	Module 1 Unit 1and 2
3	Module 1 Unit 3 and 4
4	Module 1 Unit 5
5	Module 2 Unit 1 and 2
6	Module 2 Unit 3 and 4
7	Module 2 Unit 5

8	Module 3 Unit 1 and 2
9	Module 3 Unit 3 and 4
10	Module 3 Unit 5
11	Module 4 Unit 1 and 2
12	Module 4 Unit 3 and 4
13	Module 4 Unit 5
14	Revision and response to questionnaire
15	Examination

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

Table 2: Required Minimum Hours of Study

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

ASSESSMENT

Table 3 presents the mode you will be assessed.

Table 3: Assessment

S/N	**Method of Assessment**	**Score (%)**
3	Tutor Mark Assignments	30
4	Final Examination	100
Total	100	

ASSIGNMENTS

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

EXAMINATION

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

HOW TO GET THE MOST FROM THE COURSE

To get the most in this course, you:

- Need a personal laptop. The use of mobile phone only may not give you the desirable environment to work.
- Need regular and stable internet.
- Need to install the recommended software.
- Must work through the course step by step starting with the programme orientation.
- Must not plagiarise or impersonate. These are serious offences that could terminate your studentship. Plagiarism check will be used to run all your submissions.
- Must do all the assessments following given instructions.
- Must create time daily to attend to your study.

FACILITATION

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

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

For the asynchronous facilitation, your facilitator will:

- Present the theme for the week.
- Direct and summarise forum discussions.
- Coordinate activities in the platform.
- Score and grade activities when need be.
- Support you to learn. In this regard personal mails may be sent.

- Send you videos and audio lectures, and podcasts if need be.

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

LEARNER SUPPORT

You will receive the following support:

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

COURSE INFORMATION

Course Code:	STT102 INTRODUCTORY STATISTICS
Credit Unit:	3 units
Course Status:	Compulsory

Course Blub: This is basic course designed for students with no previous knowledge of statistics. It is intended to initiate students to the subject at an introductory stage so that they acquire fundamental concept required in gaining deeper understanding of social, health and related issues.

Semester:	Second Semester
Course Duration:	13 Weeks
Required Hours for Study:	91 hours

MAIN COURSE

CONTENTS **PAGE**

MODULE 1 STATISTICAL METHOD

Module Introduction

Statistical methods are mathematical formulas, models, and techniques that are used in statistical analysis of raw research data. The application of statistical methods extracts information from research data and provides different ways to assess the robustness of research outputs.

UNIT 1 AIMS OF THE STATISTICAL METHOD

Unit Structure

1.1 Introduction

This unit focuses mainly on the aims and application of the statistical method in nursing education and practice. It gives definitions of basic statistical concepts and states salient reasons why information or data relating to health issues require statistical treatment. We thus have to look at what you have to learn in this unit in the objectives stated hereunder.

1.2 Intended Learning Outcomes (ILOs)

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

- Understand the aims of statistical techniques in nursing education and practice.
- Define and employ basic statistical terms.

1.3 Main Content

1.3.1 Definition of Statistics

Statistics is the science that deals with collecting and summarizing facts which are expressible in numerical form. It also involves the measurement and comparison of facts ultimately leading to the discovery of the existence of significant relationships between them. This is helpful in revealing trends so that important estimates or forecasts may be carried out.

You will therefore observe from the above definition of Statistics that we often ask so many questions that tend to have Statistical implications. For example: What is the average height of ten-year-old girls in Nigeria? How many Nigerian mothers wean their babies at nine months old? What is the life expectancy of a Nigerian woman? You often engage in such discussions locally or at your work-place. There is a variety of opinions on such subject yet no tangible information results except reliable data are available.

However, these questions are not only of grave consequence to the well-being of individuals and groups of people but have economics, political and health implications for the country. You will therefore notice its relevance and importance in the next section and the course in general.

1.3.2 Decision-Making

Let us reflect over the questions raised in the previous section of the unit and several everyday actions we engage in. you will notice that we tend to make several decisions in our daily life. Some of these decisions may be simple while others are consequential and involve some degree of uncertainty. You will further notice that decisions are usually made with regard to available information given or assumed.
Quite reasonably then, numerical information is preferable since it presupposes an assessment of consequences. Statistics may therefore be

generally defined as that part of decision making which relates to numerical information. Therefore, there is the need for you to be conversant with some basic concepts and terminologies presented I n the remaining part of this unit.

1.3.3 Population and Sample

Population is a collection of all individuals, items or data under consideration in a statistical study.
You can employ the term to refer to a collection of:
(i) Citizens of a country
(ii) Students in a school of nursing
(iii) Patients in a hospital ward
(iv) Animals in a zoo; and
(v) Inanimate objects e.g., cars, houses, dogs.

A population may be finite, infinite, countable or uncountable. All the unit of the population may not be reached at times, we then refer to that part of the population from which information is collected as a sample.

A sample may be random or purposive according as each sample is equally likely to be selected or not. For example, a random sample is obtained by tossing a coin, throwing a die or drawing slips from a box.
You also need to be conversant with the following terms:

1.3.4 Variable and Observation

A characteristic possessed by the numbers of a population is said to be a *variable.* A variable may take integral or real values. For instance, age, weight and height are variables. Variables may be referred to as *discrete* or *continuous* depending on whether they take specific values or not.

The size of a family is a discrete variable whereas the heights of twenty-five students of the national Open University form a set of continuous variables. Finally, we define an observation as a variable for a member of a population.

1.4 Self-Assessment Exercise(s)

Construct a table indicating the number and ages of the following category of staff in your hospital or unit:
(i) Nursing officers and above
(ii) Ward sisters and charge nurses
(iii) Staff nurses

(iv) Student and pupil nurses
(v) Unqualified staff e.g., ward underlies and nursing assistants.

1.5 Conclusion

In this unit, you have been able to identify the aims of the statistical method. You should also be able to define, recognize some basic statistical terms, and be able to relate these concepts to some case studies occurring in nursing practice and other cognate areas.

1.6 Summary

What you have been able to learn in this unit deals with the extent statistical method is required in the analysis and interpretation of figures relating to health issues. You also learned that the analysis and interpretation of figures are influenced by factors which include for instance variability of human beings in their illnesses, their reactions to them and ultimately in the treatment of these illnesses.

You also understood that the requirement of a large number of data is not essential in most cases since a smaller number of data may furnish vital information on the issue under study. The study units that follow build upon this introduction.

1.7 References/Further Readings

Jefferies, P.M.(1995) *Mathematics in Nursing,* 4th edition, Bailliere Tindall, London: Cassell and Collier, Macmillan Publishers Ltd.

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

UNIT 2 COLLECTION OF DATA: SAMPLING

Unit Structure

2.1 Introduction

This unit is concerned not only with how to plan and conduct a statistical study but with the methods of gathering adequate and reliable amount of information which needs to be treated scientifically.

It examines critically various methods for acquiring information which reflect as closely as possible the relevant characteristics of the population under consideration. The unit further guides you on how to ascertain sources of information so as to be able to determine what the study is attempting to portray. Let us look at what you should learn in this unit as stated in the objectives below:

2.2 Intended Learning Outcomes (ILOs)

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

- Plan and conduct a statistical study involving health issues
- Employ an adequate or reliable method of sampling in your statistical study
- Determine vital sources of information for your study
- Distinguish between random and non-random sampling schemes by discussing the meaning, advantages and disadvantages of each.

2.3 Main Content

2.3.1 Is a Statistical Study Necessary?

Often, it is necessary to obtain information about a subject of interest. This is usually conducted through a statistical study whose purpose is to investigate the existence of relationships between two variables or characteristics. For example, medical researchers may require information on the relationships between smoking habits and lung cancer in one situation or vasectomies and incidence of prostate cancer in another case. Both cases require statistical methodology.

Statistical methodology is divided into two main branches- *descriptive* and *inferential statistics.* In its own case, *descriptive statistics* is concerned with the organization, presentation and summarization of data. Techniques of descriptive statistics usually entails the presentation of data in tables and graphs as well as summarization of a collection of data by means of numerical values like the average, median, range and variance.

Inferential statistics on the other hand involves procedures employed to draw conclusions about a large body of data called a *population*. The population in a collection of data is usually based on a smaller set of data referred to as *sample.* We shall discuss the techniques of both descriptive and inferential statistics in subsequent study units.

To obtain information on any subject of statistical interest, you will therefore have to conduct and plan a study. If information required is not already available from a previous study, you will plan a new study by acquiring information through selection of a sample or census taking.

Sampling is the more commonplace method to gather information. If you consider sampling suitable, you then decide on what sample of the population is the most representative. In the remaining part of this study unit, you will various sampling techniques.

2.3.1 Simple Random Sampling

A simple random sampling process is one for which is possible sample is equally likely to be selected. A sample obtained by simple random sampling is said to be a *simple random sample.*
You should note that they are two types of simple random sampling. One is *simple random sampling with replacement,* where a number of

the population is selected more than once. The other is *simple random sampling without replacement*, where a member of the population is selected at most once.

You should also note that obtaining a simple random sample by picking slips of paper out of a container is not practical, especially when the population being sampled is large. Several practical processes for getting simple random samples exist. A table of random numbers is one common method. You should also note that computers can be employed to obtain a simple random sample from a population.

A good choice of a computer package to obtain a simple random sample from a population is Minitab. The appropriate command for this is referred to as Sample.

Although the simple random sampling is the easiest to understand, yet it has some shortcomings. These include its failure to provide sufficient coverage when information about sub-populations is required. Another drawback is its impracticality when the members of the population are not contiguous. In the remaining part of the unit, we will examine some other commonly employed sampling processes that are more suitable than simple random sampling.

2.3.2 Systematic Random Sampling

Systematic Random Sampling is easier to implement than simple random sampling. The process consists of the following three steps:

(i) The population size is divided by the sample size and the result is rounded up to the nearest whole number k say,

(ii) Obtain a number a, say, from a random number table (or from a simple device)

(iii) A must lie between *I* and *n.*

(iv) Select for the sample $a, a + k, a + 2k,$ which are members of the population.

(v)

You will observe from the explanation above (i.e., in 2.4) that systematic random sampling provides comparable results with simple random sampling except that listing of the members of the population has a cyclical pattern (e.g., male, female, male).

2.3.3 Cluster Sampling

This method is useful in the case when members of the population under consideration are widely scattered geographically.
Cluster sampling can be executed in the following three steps:

(i) The population is divided into groups (clusters);
(ii) A simple random sample of the clusters is obtained; and
(iii) The sample then is then taken as all of the members of the clusters obtained in (ii). Cluster sampling has a advantage in that each cluster does not mirror the entire population. Members of a cluster are frequently more homogeneous than the members of the population as a whole.

2.3.4 Stratified Sampling

This sampling method is often more reliable than cluster sampling. It is implemented by dividing the population into sub-populations (or strata) and then sampling is done from each stratum. Stratified sampling has the advantage that each stratum should be homogeneous related to the characteristic under consideration.

Let us now consider the last method of sampling in this course.

2.3.5 Multistage Sampling

Multistage sampling (as its name suggests) is a combination of the earlier mentioned sampling processes. It is frequently used by government agencies. For example, multistage sampling may be conducted by the Federal Ministry of Health on the population to acquire information on illnesses, injuries, mortality rate and other health issues.

Let us now consider examples to illustrate some of the sampling procedures discussed.

Example 2.2.1

The annual salaries for five Government Officials are as shown in Table 2.3.1 below. These are in millions of Naira, rounded up to the nearest million.

Official	**Salary**
Governor (G)	5
Deputy Governor	4
Secretary to the State Govt. (SSG)	3
Head of Service (H)	2
Permanent Secretary (P)	1

Table 2.2.1: *Annual Salaries for five Government Officials.*

(i) List the possible samples (without replacement) of two salaries that can be obtained from the population of five salaries.
(ii) Describe a method for obtaining a simple random sample of two salaries from the population of five salaries.
(iii) State the chances that any particular sample of two salaries will be the one selected for the sampling procedures described in (ii).

Answer

(i) You will observe that there are ten possible samples of two salaries from the population of five salaries. The listing in the table below is done using the letters in parenthesis to represent the officials.

	Officials Selected	***Sample Obtained***
G, DG	5, 4	
G, SSG	5, 3	
G,H	5, 2	
G,P	5, 1	
DG, SSG	4, 3	
DG, H	4, 2	
DG,P	4, 1	
SSG, H	3, 2	
SSG, P	3, 1	
H, P	2, 1	

Table 2.2.2 *possible samples of two salaries from the population of five samples*

(i) Write each letter corresponding to the five officials 0n separate pieces of paper. Place the five slips of paper in a box and shake the box. You will have to pick two of the slips of paper while blindfolded.

(ii) You will notice that the procedure used in (ii) is a simple random sampling. You will also notice that each of the possible samples of two salaries is equally likely to be selected. These are ten possible samples and the chance of selecting any particular sample are 1/10 (1 in 10).

Example 2.2.2

A study that was conducted by a health inspector requires the determination of the extent of vaccination against poliomyelitis amongst five-year primary school pupils in a Nigerian city. He obtained a list of all currently enrolled first year pupils with their names being listed consecutively. A set of 1000 random numbers were obtained from a

table of random numbers. Pupils on the list whose numbers corresponded to the random numbers obtained from the table were selected for inclusion in the study. Determine what type of sample selection scheme was used. Generalize the findings to a larger group.

Answer

You will notice that a simple random sampling process was used. You will also observe that the use of this process was possible because a listing of all pupils in the city was possible. Investigator-induced bias was impossible since the names of the pupils in the study were chosen through the use if a random number table.

It is not possible to generalize to a larger group of pupils in the nation because this is not statistically justifiable and subjective arguments about the bias of the population sampled come to play.

Example 2.2.3

Is the use of cluster sampling possible in Example 3.3.2? if so, explain how this may be employed

Answer

This is possible. The health inspector would select a simple random sample of schools from a list of all primary schools in the Nigerian city. He would then visit only schools included in the study sampled where he should either interview all first-year pupils or a simple random sample of first year pupils. You will notice that the cluster or primary sampling unit is the primary School while the pupil is the Second sampling unit.

Example 2.2.4

A Nigerian town has 25 major homesteads with each divided into 150 houses. A health inspector wishes to carry out a survey on the attitudes of the two dwellers toward HIV. Discuss possible sampling schemes for carrying out the survey.

Answer

(i) A simple random sample of names drawn from a list of all town dwellers is not practicable list such list does not exist or is not up-to-date.
(ii) A simple random sample of homesteads from among the 25 homesteads is taken.

Next, a simple random sample of houses is then selected. The health inspector now takes a random selection of even or odd numbered houses. You will notice that this sample selection procedure involves methods of simple random samplings, cluster sampling and systematic sampling.

2.4 Self-Assessment Exercise(s)

A new nursing care procedure for hospitalized diabetic patients was instituted using diabetic patients from hospitals randomly selected among all hospitals in Nigerian city. Discuss the type of sampling scheme used and to what extent can the results be generalized.

2.5 Conclusion

In this unit, you have learned that nursing practice and research involves the conduct of statistical survey dealing with observations on humans. This type of research or practice is non-experimental in nature because it is not always possible to manipulate the study environment.

You also learned that statistical surveys in nursing practice and research fall under three categories, namely, the *cross-sectional study, retrospective study and the prospective.* You learned that in general, these studies are often conducted for the purpose of establishing associations between two disease states and presence of a risk factor.

You also learned that the nature of these surveys determines the procedure of collection of data or information required for them. Furthermore, you learned the various procedures for the collection of data and what particular sampling process is suitable for conducting any of the surveys.

2.6 Summary

In an investigation to determine the different arm positions on blood pressure determinations, *twenty-five* patients were available on a general clinical research unit. One group had their blood pressures taken using the standard arm position; the second group held their arms to their sides while blood pressure readings were made. Discuss what type of

sampling scheme would be used to ensure proportionate representation for both groups of patients.

2.7 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

Weiss, A.L (1996) *Elementary Statistics*, 3rd edition, Addison-Wesley Publishing Co. Inc.

UNIT 3 COLLECTION OF DATA: BIAS

Unit Structure

3.0 Introduction

In the previous unit, attention was devoted mainly to the situation in which a sample of observations could be deliberately drawn for study from some known population. You further saw that the fundamental issues were then to define and determine a means of drawing a random sample from the population.

You need to be aware however, that in medicine or nursing practice, the situation is different. This is because medical practitioners, nurses and professionals in allied disciplines have to accept whatever sample of observations that present itself in the natural cause of daily events.
In this case, you will notice that professionals in the health sectors must consider very carefully whether the sample is representative of all patients and not in any way biased.

You will learn in this unit how bias may be introduced into any statistical survey conducted in the health sector.

3.1 Intended Learning Outcomes (ILOs)

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

- identify bias in any statistical survey
- recognizes sources of bias in a statistical survey
- describe and recognize a sample which is representative of the population to which it belongs.

3.2 Main Content

3.2 What is Bias

In a statistical survey, a sample drawn from a population of observation is said to be biased if the sample is not representative of the parent population of which it is part.

Bias may be investigator-induced, that is, when it is deliberate or it may be unforeseen or unrealized. In the case when it is deliberate there exists a lack of comparability between the sample and the population. For example, if the treatment of malaria by therapeutic means were confined to those patients who cough, then it is clear that these patients are not representative of the sample of all patients who suffer from malaria.

To compare their subsequent times of recovery with that of all patients is statistically unjustifiable. However, if the bias is unrealized or unforeseen, it is impossible for the sample to be representative and the bias could have resulted from the method of collection of the observations or to the limited field in which the survey was conducted.
In either case, interpretation of the statistical procedures will be faulty.

You need to be aware that a series of examples in medicine, nursing practice and research contribute to bias of samples of observations. We will consider several sources of this bias in the remaining part of the unit.

3.2.1 Examples of Bias

In medicine as well as in nursing practice, it is seldom possible to draw a sample of observations for study from some known population. The situation is such that a sample of observations has to be accepted from sources that present themselves in the natural courses of daily events.

These sources are guided by so many unavoidable factors whose possible presence cannot be too carefully remembered or taken into account in the eventual interpretation of all the statistics. We now mention some of these sources.

3.2.2 Sex Ratio at Birth

You will find out that the frequency with which male and female births are recorded in the hospitals and local government councils is unlikely to be representative of the births in the country as a whole. This important information is distorted by scanty records in the rural areas or by cultural preferences like proud parents who are likely to record only the births of their sons.

Whatever the reasons, sex ratio at birth may not be easily available and as such one cannot generalize from such a sample of births about the population of the whole country without entertaining some degree of bias.

3.2.3 Hospital Statistics

Another factor contributing to bias in statistical survey in medicine and nursing practice is hospital statistics. For example, hospital statistics of a disease can very rarely be regarded as representative of all cases of that disease. This is because patients are often drawn from certain areas which may have differing populations in age, sex and ethnicity. They may also come from particular social classes. It is also observed that in many diseases only those patients who are seriously ill are likely to be taken to hospitals.

You will then observe from the above points that it is obvious that it is not possible to generalize the fatality rate or success of treatment of some diseases with any approach to accuracy of samples of the population of hospitals statistics without incorporating a measure of bias.

3.2.4 Treatment Day

One other source of statistical difficulty with the value of some form of treatment of diseases is the treatment day. For instance, the level of fatality rate or success of treatment of a disease at different stages is likely to be seriously biased.

This is because not all patients suffering from a particular disease go to the hospital at the same stage of the disease. Some may go to the hospital at the onset of the disease; others may be taken to hospital when their condition has become serious or at the point of death.

3.2.5 Statistics Relating to Post-mortem

In an attempt to obtain more accurate data, emphasis is usually placed upon post=mortem statistics. However, increased accuracy is gained at the expense of employing material that may be highly selective.

It is rather unlikely that every death would be subjected to autopsy while those chosen may not be chosen randomly.

You should notice from the above points that any feature observed at death is most likely to be representative of the living population.

3.2.6 Patients Follow-up studies

Incomplete follow-up studies of patients are frequently subject to selective influences. If you base your conclusions upon patients who are successfully followed up after hospital admission or otherwise by assuming that the results recorded for such a group are unchanged if lost-to-sight patients are not added, then your results are biased.

You need to make the follow-up comprehensive to avoid bias of your results.

3.2.7 Infant Feeding

Statistical comparisons of breast or artificial feeding of infants are likely to be influenced by selective factors. This is due to what measure of the degree of its advantages is considered.

Other factors include how random is the sample of the population considered and at what stage is the mode of feeding introduced. You will need to consider these factors very closely before interpreting your results.

3.2.8 Use of Questionnaires

Here we look at another source of bias in a sample of population in medical in nursing practice. This involves inquiries carried out by means of questionnaires. You should be able to recognize that such inquiries, replies to the questions put are received from only a proportion of the individuals to whom the form is sent.

There can never be the slightest guarantee that the individuals who decide to reply are representative of the sample of all the individuals approached. You may correct the situation by stating the number of

missing questionnaires or items and take this into consideration in your interpretation of your results.

3.3 Self-Assessment Exercise(s)

Define and give five examples or sources of bias in medical or nursing practice.

3.4 Conclusion

In this unit, you have learned that a statistical study may be influenced by some degree of bias. You have also learned some sources of this bias in medical and nursing practice. In addition, you also learned the various ways in which these sources of bias could be avoided or corrected.

3.5 Summary

In this unit, you learned that if you wish to generalize from some sample group of observations, you must determine a sample which is representative of the population to which it belongs.

If you select or accept samples deliberately, you should realize that bias may occur through the operation of several factors leading to a sample which is not representative of the total population. Self-selection of members of a group, absence of some of the required records e.g., by individuals who do not reply to a questionnaire are common forms of bias.

When you generalize from a sample or make comparisons, you should be aware of the possible presence of such biases.

3.6 References/Further Readings

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics*, 12th edition, E.Arnold, London.

Knapp, R.G (1985). *Basic Statistics for nurses*, 2nd edition, Delmar Publishers Inc. N.Y.

Jefferies, P.M.(1995) *Mathematics in Nursing*, 4th edition, Bailliere Tindall, London: Cassell and Collier, Macmillan Publishers Ltd.

UNIT 4 COLLECTION OF DATA: FORMS OF RECORD

Unit Structure

4.0 Introduction

In this unit, we will discuss the last factor that influences the collection of data when conducting a statistical survey. This deals with forms of record.

You should be aware that in all scientific work, we are involves in asking questions. In nursing practice and research, for instance, you may wish to determine the effects of a specific treatment employed upon patients suffering from a specific disease. Whatever mode of inquiry you adopt, you are often asking questions.

It is therefore desirable that you need a form of record of these questions and their associated answers. In doing so, you need to construct the form in such a way as to include clear and definite questions. You should also anticipate what answers would be given and how the answers can be analyzed and set into a statistical table at the end of the survey. We now look at the objectives of this study unit.

4.1 Intended Learning Outcomes (ILOs)

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

- Construct forms of record for collection of data
- Identify and present unambiguous questions in a form of record for collection of data.
- Set up a statistical table from a form of record.

4.2 Main Content

4.2.1 Questions and Answers

To formulate questions or headings for inclusion on a form of record, you need to bear in mind the following important principles:

(i) Consider if there is any ambiguity in the questions and consequently in the responses received.

For example, you need to construct the form in such a way that age last birthday is required. You will be aware that date of birth may even be preferable to the age last birth since date of birth is constant while age can be calculated from it from time to time.

(ii) You need to ensure that every question in the form is self-explanatory. Respondents need not consult a separate sheet of instructions to answer question on the form of record.

(iii) You should ensure that every question requires some answers. This guarantees that the respondents offer useful information or possesses a characteristic sought for.

(iv) You should specify the degree of accuracy required in answering every question. For example, if body temperatures are to be taken orally or rectum.

(v) You need to ensure that any form of record which must be completed by many people should be worded simply and logically.

- You should ask questions that vary widely on circumstances.
- You should distribute a large number of questions over different samples.
- You should be aware that a shorter form with questions may promote greater accuracy of reply and also reduce the amount of non-response.

4.2.2 Design of Form of Computer Use

Nowadays, almost all that are transferred to computer files are usually by typing the information in, by hand from the completed questionnaires. You should ensure that the form be designed so as to facilitate this transfer.

4.3 Self-Assessment Exercise(s)

Construct a form of record for collection of data in a hypothetical case involving the treatment of patients suffering from tuberculosis.

4.4 Conclusion

In this unit, you have learned the fundamental principles of constructing forms of record involve in constructing statistical surveys.

4.5 Summary

One of the most decisive and difficult tasks in any you learned that questions posed should be clear answer and entail a standard of accuracy.

You also learned that the forms need be designed so inquiry is the construction of a suitable form of record and unambiguous. Each question should require some as to facilitate transfer of information to a computer file.

4.6 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

Weiss, A.L (1996) *Elementary Statistics,* 3rd edition, Addison-Wesley Publishing Co. Inc.

UNIT 5 PRESENTATION OF DATA

Unit Structure

5.0 Introduction

This unit is concerned with the meaningful manner in which raw data (which are usually in the form of large sets of unorganized numerical values) are summarized and interpreted so that important features and trends may be identified.
A set of data may be presented in tables or described by means of diagrams, charts and graphs. Before discussion these terms, let us look at what you should learn in this unit as stated in the objectives below:

5.1 Intended Learning Outcomes (ILOs)

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

- Tabulate and organize raw data;
- Present data by means of tables, diagrams, charts and graphs and
- Interpret and highlight important features of the data through the tables, diagrams, charts and graphs.

5.2 Main Content

5.2.1 Tabulation

The purpose of classifying and tabulating a set of raw data is to simplify and facilitate the interpretation between trends and features arising from statistical investigation. This is done by means of tables which are of three types namely, *source (or reference) tables, working and summary tables.* Further analysis is based on the source table while a working table is one on which initial calculation are carried out. A summary table in its own case is usually found in books so as to facilitate reference.

You need to be aware that a statistical table should have the following characteristics:

(i) A general title indicating the purpose of the table;
(ii) A column title indicating the order of classification along the columns;
(iii) A row title indicating the order of classification along a row;
(iv) Source of information contain in the table usually indicated at the bottom; and
(v) Units of data in the table.

5.2.2 Diagrams, Charts and Graphs

After tabulating a collection of data, pictorial and diagrammatic description is carried out. This is to convey information and you will learn the simplest of these methods in the next part of the unit.

5.2.3 Pictograms.

The simplest method of presentation of information is through pictorial figures. These are commonly used in advertisement and in the print media. They convey broad relationships involving two or more variables. Let us now discuss and illustrate each type of such figures in the following part of this study unit.

Example 5.1

Construct a diagram to feature the number of nurses in General Hospitals in the country in 1990/1991 and 2000/2001 using the following data;

Time	*No. of Nurses (in thousand) in General Hospitals*
1990/1991	20
2000/2001	45

Answer

The pictogram is drawn using a related picture of a human being since the data collected are population. One picture stands for ten thousand nurses, hence

20 thousand have 2.0 pictures

45 thousand have 4.5 pictures

1990/1991

2000/2001

Figure 3.1 Number of nurses in General Hospitals in 1990/1991 and 2000/2001.

5.2.4 Block and Bar Diagrams

A bar diagram or chart consist of a line, bar or columns of equal thickness but variable lengths. The bars may be vertical or horizontal. You will understand this concept from the example given below;

Example 5.2

Admission of students into a nursing school in Nigeria for ten successful years is given in the table below:

Year	*No. of students (in hundreds)*
1990	400
1991	700
1992	750
1993	800
1994	650
1995	600
1996	1000
1997	1200
1998	1300
1999	1400

Table 5.1 Construct a bar chart for the data

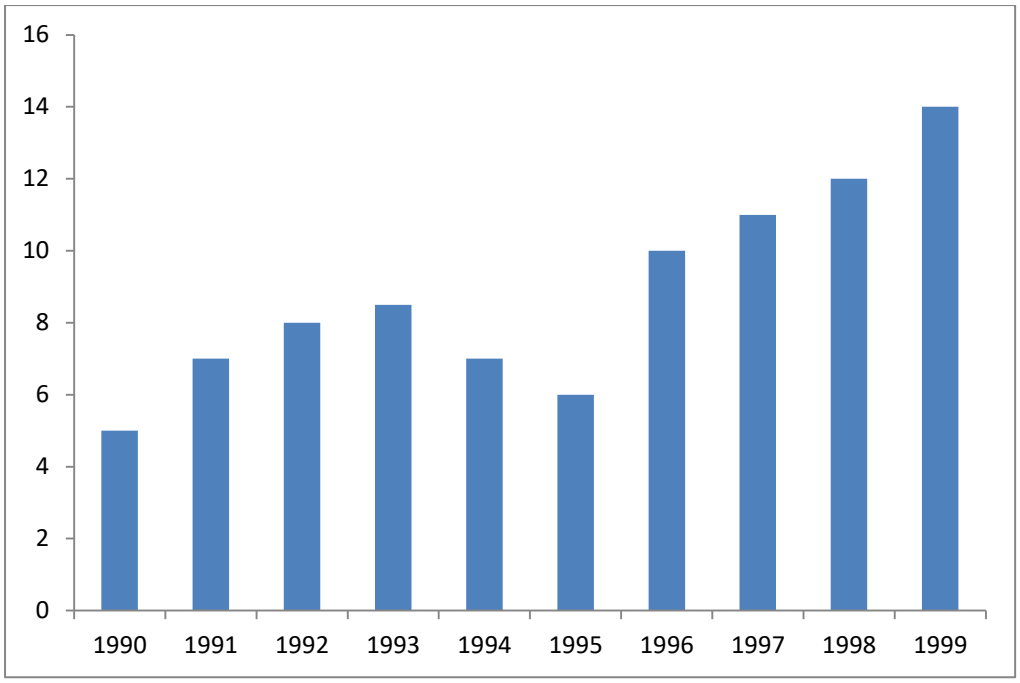


Figure 5.2: Bar Diagram; Admission of Nursing Students in Nigeria 1990-1991

5.2.5 Pie Chart

A pie chart is a circle divided into pie-shaped pieces proportional to its relative size. We shall use the next example to describe the steps involve in constructing a pie chart.

Example 5.3
Attitude scores of five newly admitted nursing students towards alcoholic patients are given as follows;
The scores are determined by
(1= very negative, 2= slightly negative, 3= slightly positive, 4= positive, 5= very positive)
5,2,3,1,4
Use a pie chart to convey the information

	Attitude	***Relative size***	***Percentage (%)***	***Degrees***
	5	0.33	33.0	119
	4	0.27	27.0	97
	3	0.20	20.0	72
	2	0.13	13.0	47
	1	0.07	7.0	25
Total	**15**	**1.00**	**100**	**360**

Step 1: You will calculate the percentage that each attitude score is of the total. These are shown in the table above

Step 2: you will find the size in degrees of each proportion in step 1. The total is 360 which is the angle in the circle.

Step 3: You will draw a circle of reasonable size and mark the angles obtained in step 2.

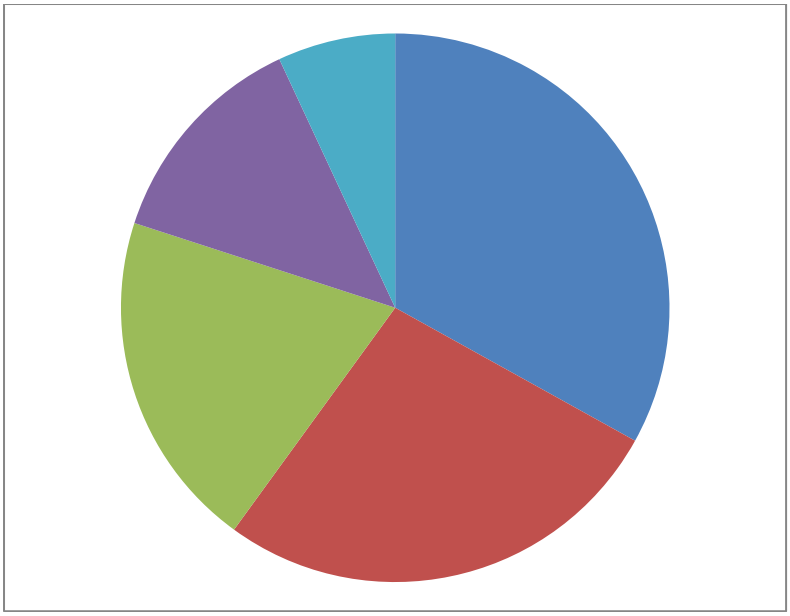


Figure 5.3 Pie chart: Attitude score of 5 newly admitted nursing students towards alcoholic patients.

Graphs

A graph is a pictorial representation which shows the relationship between a variable of interest (Dependent variable) and that on which this variable of interest depends (independent variable).

A very good graph should possess a clear layout and indicate the following:

(i) Title
(ii) Unit of measure
(iii) Scale
(iv) Source of data

The following example will teach you how to construct good graph.

Example 5.4

Construct a good graph for the data presented in Example 5.3

The scale should usually start from zero. You need to note that the position on the graph is located by the coordinates of the point, for example, the position of the number of students admitted in 1999 given as 1400 is located by moving the x-axis until 1999 and at this point, you will move upward on the scale until you get to the value as shown in figure 5.2 when you have located all the points in this way, then join them in the order presented in the table by means of lines.

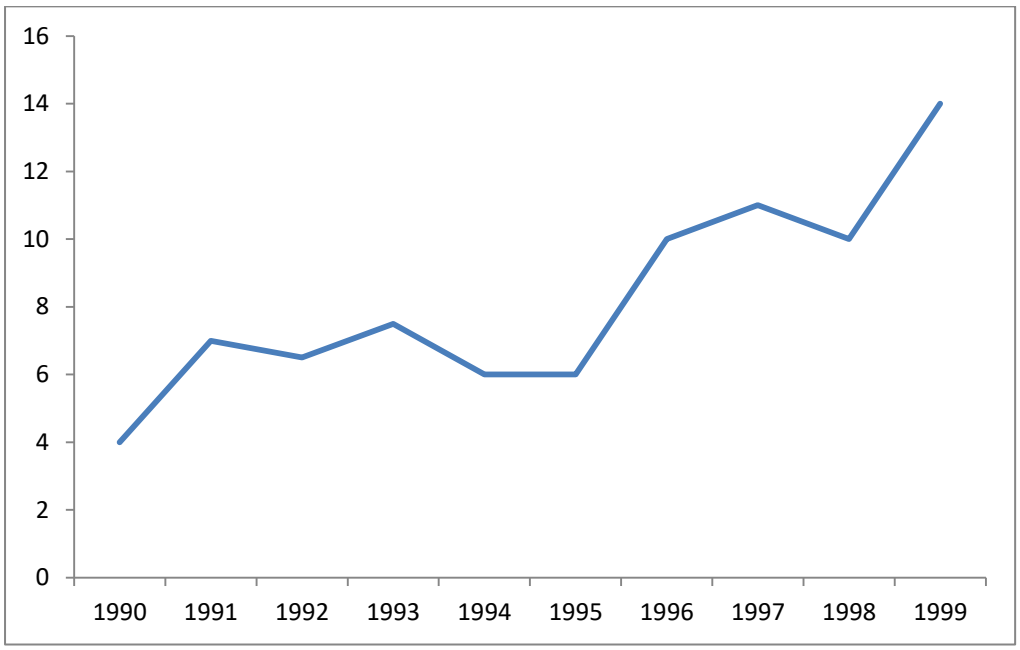


Figure 5.4 A line graph: Admission of Nursing Students in Nigeria 1990-1999

5.3 Self-Assessment Exercise(s)

Of 100 patients in an orthopedic hospital who were asked for their room preferences, 50 wanted private rooms, 40 wanted semi-private and 10 would make do with any room. Present this data by means of a bar chart.

5.4 Conclusion

In this unit, you learned how raw data are summarized and interpreted. In this wise, you saw that for a comprehension of a series of figures tabulation is essential.

Furthermore, you saw that data may also be presented as graphical representations and charts. There is need for these representations to be

simple, perfectly clear and self-explanatory. You need to know that conclusions could be drawn from graphs only after careful consideration of the scales adopted.

5.5 Summary

The following critical concepts learned in this unit are:

- The raw data resulting from a sample survey or a census are usually in the form of unorganized values.
- A set of data must be organized and summarized so that important features are apparent.
- A set of data may be described pictorially by means of diagrams, bar charts and graphs.

5.6 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses*, 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics*, 12th edition, E.Arnold, London.

Weiss, A.L (1996) *Elementary Statistics*, 3rd edition, Addison-Wesley Publishing Co. Inc.

MODULE 2 MEASURE OF LOCATION AND DISPERSION

Module Introduction

A measure of central location is the single value that best represents a characteristic such as age or height of a group of persons. A measure of dispersion quantifies how much persons in the group vary from each other and from our measure of central location. Measures of location are statistical methods used to show representative value in the data set. Mean, mode and median are examples of commonly used measures of location. Measures of dispersions are methods used to measure variability or spread of data in the data set. They include; variance and standard deviation.

Unit 1 Frequency Distribution
Unit 2 Cumulative Distribution
Unit 3 Measures of Location
Unit 4 Measure of Dispersion
Unit 5 Correlation

UNIT 1 FREQUENCY DISTRIBUTION

Unit Structure

1.0 Introduction

In this unit, you will learn one of the most significant and fundamental process in which raw and un-organized data are displayed. This involves

tabulating the number of times each score in a collection of data occurs in the sample. This number is referred to as the *frequency* of each score and the table associated with this display is called the *frequency distribution.*

You will learn how this table can be constructed for discrete and continuous data. You will also learn how to compare the frequency distributions of two different sets data in this study unit. This comparison for ease of interpretation is usually done through the proportion or percentage of the total number of observations falling into each interval and is known as the *relative frequency* i.e.
Absolute frequency relative frequency = 100%

Total number of observations
You will also study graphical representation of either frequency or relative frequency tables and histograms. The concluding portion of the unit discusses another important graphical representation of a frequency table. This is the frequency polygon. We will first examine what you will learn in this unit in the objectives stated hereunder:

1.1 Intended Learning Outcomes (ILOs)

By the end of this unit, you should be able to:
- Define the following terms

a. Frequency distribution; and
b. Relative frequency

- Construct absolute and relative frequency tables;
- Present a frequency table by means of a histogram;
- Identify shapes of frequency distribution as symmetric or non-symmetric.

1.2 Main Content

1.2.1 Frequency Distribution and Frequency Tables Using Discrete Data

This concept will be explained using the attitude score of 20 cancer patients to nursing care. The patients were required to indicate their feeling about nursing care in respect of alleviation of their discomfort. The attitude score ranges from 1 to 10.
The responses are stated as follows:

8	8	6	5	2	4	6	4	6	6
5	6	6	2	8	7	6	3	2	6

You will tabulate the number of times each score appears in the simple and display the result as given below:

Table 1.1: *Frequency Distribution of Attitude Scores of Cancer Patients to Nursing Care*

Score	***Frequency***
1	0
2	3
3	1
4	2
5	2
6	8
7	1
8	3
9	0
10	0
Total	**20**

The table you have obtained is called a *frequency table* and it displays the frequency distribution of the 20 attitude scores.

This is the simplest form of frequency distribution. In the course of the unit, you will learn how this concept may be expanded to involve the use of intervals of data instead of individual scores.

We will construct another frequency table to illustrate the concept more concisely. This is done using examination results of 40 nursing students in the final degree examination at a local university. The scores are as follows:

40	25	30	30	40	40	41	40	42	40
40	40	42	42	50	52	55	60	60	65
40	61	62	60	72	60	40	42	48	42
70	67	65	40	42	42	40	40	40	40

Table 1.2: *Frequency Distribution of Examination Results of 40 Nursing Students*

Score	*Frequency*
25	1
30	2
40	14
41	1
42	7
48	1
50	1
52	1
55	1
60	4
61	1
62	1
65	2
67	1
70	1
72	1
Total	**40**

It is also possible to represent the table data in Table 6.1 and 6.2 by means of frequency diagram as shown below, you will observe that in this diagram frequency is given on the vertical axis while the collection of data is given on the horizontal axis. The height of each bar stands for the frequency of occurrence of each score.

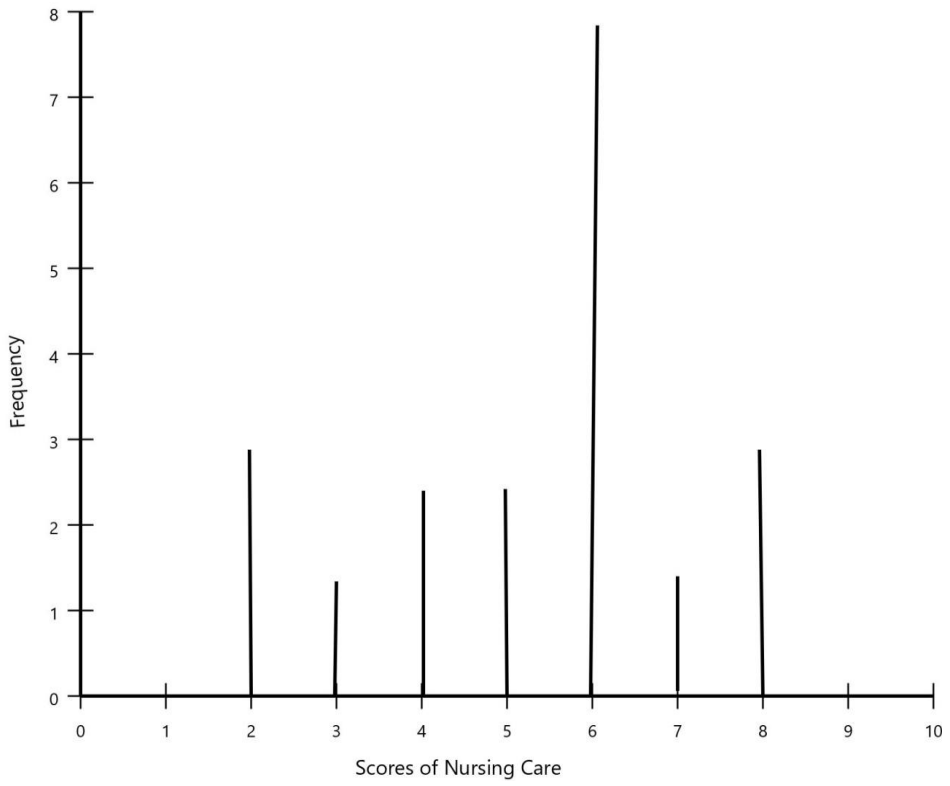


Figure 1.1: Frequency line graph of Attitude Scores of cancer patient to Nursing Care

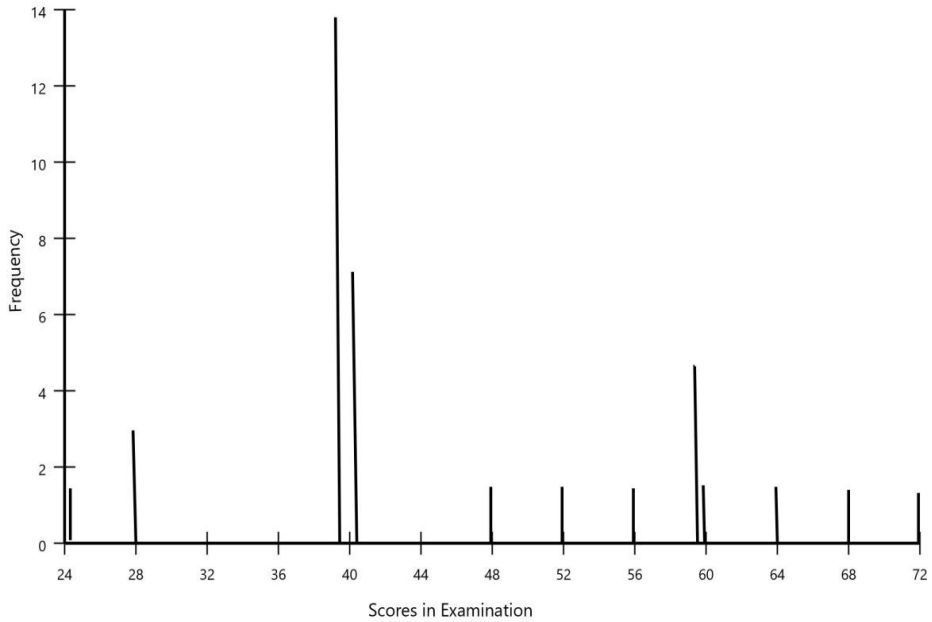


Figure 1.2: *Line graph of Scores in Exam and Frequency.*

One other important feature you will study is the frequency diagram or the shape of distribution of a collection of data. Let us consider the examples shown in figure 6.3 below.

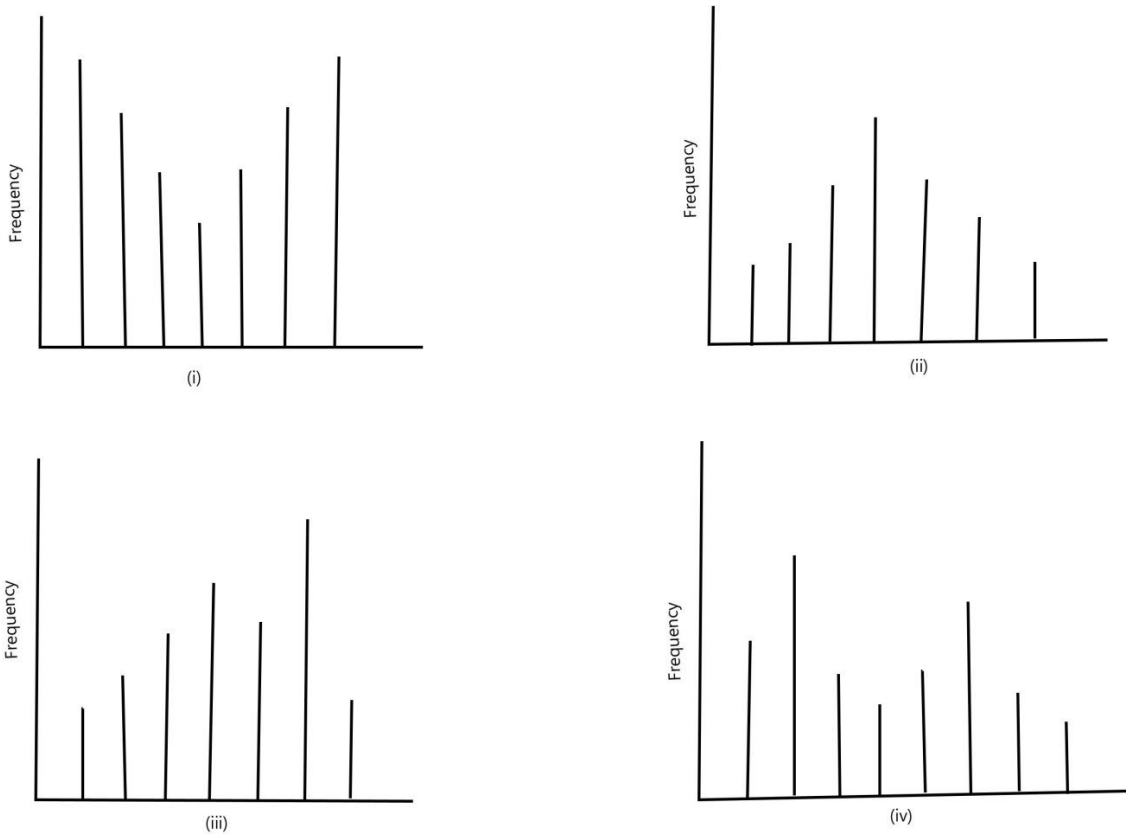


Figure 1.3: *Symmetric and Non-Symmetric Frequency Distribution*

You will observe that in (i) and (ii) in figure 6.3; the frequencies are almost equal on either side of a center point. We refer to a distribution of data values with this shape as a symmetric distribution. On the other hand, in (iii) and (iv) in the figure, the frequencies are not equal on either side of the center point. Such a distribution of data values with this shape is said to be *non-symmetric.*

1.2.2 Frequency Tables Using Class Intervals

When the number of data values is small, it is meaningless to record the frequencies associated with individual scores in the sample. However, when a large sample of data is under consideration, the number of distinct sample data may be too large to be meaningful.

In this case, you will divide the data values into groups (classes or interval) and then you will record the number of data values which fall into this interval. To illustrate this technique, let us consider the data representing the weights in pounds of 30 cancer patients given below:

62.5	62.6	62.8	62.9	63.4	64.5
64.7	62.8	70.4	71.5	70.5	70.3
64.1	63.4	70.8	72.9	70.7	70.9
68.7	67.5	66.4	66.9	70.3	80.9
80.4	79.6	78.2	77.6	77.3	80.2

You will notice that many of the data values occur once or twice in the sample. Hence, tabulating the frequency of occurrence of each possible observation provides minimal information about the raw data alone.

If more meaningful information about the data is to be obtained you will summarize the data values into groups or intervals and determine the frequency with which the actual data values fall into each of the intervals.

Let us see how to construct the frequency distribution in this case. First, you will arrange the weights in ascending order of magnitude.

Table 1.3; Weights (in pounds) Arranged in Ascending Order of Magnitude

62.5	63.4	67.5	70.7	77.6
62.6	64.1	68.7	70.8	78.2
62.8	64.5	70.3	70.9	79.6
62.8	64.7	70.3	71.5	80.2
62.9	66.4	70.4	72.9	80.4
63.4	66.9	70.5	77.3	80.5

You will next determine the intervals of weight values. There is no standard rule as to finding this that usually *five to twenty* intervals will be sufficient to convey the distribution and shape of the data values.

You should note that too few intervals result in loss of valuable information about shape and distribution of data while too many intervals do not convey any meaningful information. In general, you will arrive at the right choice of number of suitable intervals by trial and error.

It is however most advisable you follow the general guidelines stated hereunder for the constructions of frequency tables:

(i) Choose *five* to *twenty* intervals
(ii) Intervals should have equal width.
(iii) You should not allow the end points of the intervals to overlap.
(iv) You should include the total number of observations in the table.
(v) You should indicate the unit of measurement (e.g., pounds, centimeters, etc.) at the beginning of the table.

Going back with the illustration with the weights of 30 cancer patients, you may choose class intervals (or class limits) of distance 1.9 to construct the frequency table below.

Table 1.4: Frequency Distribution of Weights of 30 Cancer Patients showing the class intervals of distance 1.9

Weight (Pounds)	**Frequency**	**Relative Frequency (%)**
61.0−62.9	5	16.7
63.0−64.9	5	16.7
65.0−66.9	2	6.7
67.0−68.9	2	6.7
69.0−70.9	7	23.3
71.0−72.9	2	6.7
73.0−74.9	0	0.0
75.0−76.9	0	0.0
77.0−78.9	4	13.3
79.0−80.9	3	10.0
Total	**30**	**100.1**

1.2.3 Relative Frequency

You will discover that it is of interest most times to compare the frequency distributions of two different collections of data. Let us look at a situation where we compare frequency tables showing distribution of ages of 100 girls suffering from Vesico Vaginal Fistula (VVF) with a

frequency table of ages of 1000 normal girls selected at random from the population of normal girls.

Suppose you need to compare 5-girls suffering from VVF and of ages 10-12 years with 100 normal girls of the same age group. Then, the proportion of 50 girls with VVF in the age range 10-12 years is much greater than that for the normal.

To facilitate interpretation and comparison, frequency tables have to include a column on relative frequency. This is given by
Relative frequency = Absolute frequency
Total number of observations

We shall illustrate this concept with the frequency table of weights of 30 cancer patients.

Table 1.5: *Frequency Distribution of Weight of 30 Cancer Patients*

Weight (Pounds)	**Frequency**	**Relative Frequency (%)**
61.0−62.9	5	16.7
63.0−64.9	5	16.7
65.0−66.9	2	6.7
67.0−68.9	2	6.7
69.0−70.9	7	23.3
71.0−72.9	2	6.7
73.0−74.9	0	0.0
75.0−76.9	0	0.0
77.0−78.9	4	13.3
79.0−80.9	3	10.0
Total	**30**	**100.1**

1.2.4 Histogram

Histograms are graphical representations of either frequency or relative frequency tables.
To construct histograms, you will place the true class intervals on the horizontal scale and the frequencies or the relative frequencies on the vertical scale ensuring that the height of each bar is equal to the frequency (or relative frequency) of each interval.
You must ensure that the following two points are noted in the construction of a histogram. These are that:

(i) A histogram must be self-explanatory, and have a title so that relevant aspects of the data could be identified.
(ii) The axes should be clearly labeled to reflect the scale of measurement.

Let us see how to construct a histogram using the data in Table 6.6 below:

Table 1.6: Frequency Distribution of Weights for 30 Cancer Patients.

Weight (Pounds)	*Frequency*	*Relative Frequency (%)*
60.95−62.95	5	16.7
62.95−64.95	5	16.7
64.95−66.95	2	6.7
66.95−68.95	2	6.7
68.95−70.95	7	23.3
70.95−72.95	2	6.7
72.95−74.95	0	0.0
74.95−76.95	0	0.0
76.95−78.95	4	13.3
78.95−80.95	3	10.0
Total	**30**	**100.1**

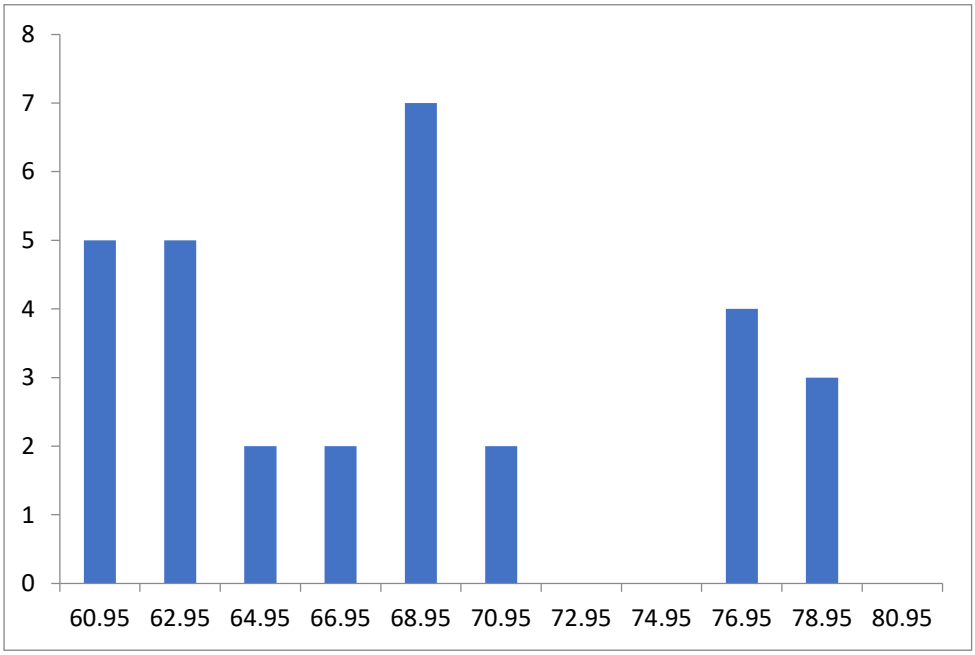


Figure 1.4: Histogram of weight (in pounds) of 30 cancer patients

1.2.5 Frequency Polygon

A frequency polygon like a histogram is a graphical representation of a frequency table. To construct a frequency polygon, you place class interval and the frequency (or relative frequency) on the horizontal and vertical axes respectively. The frequency corresponding with each interval is indicated by a dot placed above the mid-point of the interval.

You next join the dots by straight line. You should be aware that frequency polygons or relative frequency polygons are useful when comparing the frequency distributions of two or more sets of data.

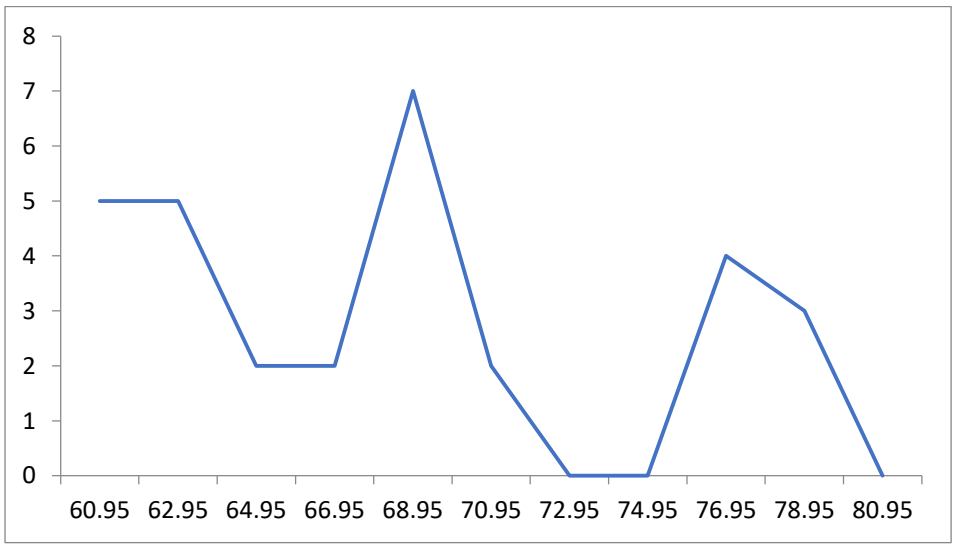


Figure 1.4: Frequency Polygon of weight (in pounds) of 30 cancer patients

1.3 Self-Assessment Exercise(s)

In a study of age distribution of patients in an orthopedic hospital, the following ages were recorded:

51	35	45	52	53	32	31	44	47	35	52
	36	44								
45	44	32	48	44	44	33	53	44	44	47
	44	44								
44	55	44	34	54	44	45	48	32	44	47
	58	50								
37	44	47	50	46	38	57	49	40	51	38

Arrange the data above in a frequency table.

1.4 Conclusion

In this unit, you have studied several graphical representations of data. These representations are used in interpreting features of data. They are descriptive in nature and we shall build up on these tabular presentations in the next part of the course.

1.5 Summary

You learned the following concepts in this unit:

- The raw data resulting from a sample survey or census are usually unorganized.
- A collection of data must be organized and summarized so as to reveal the significant features.
- A collection of data may be described by frequency tables, histograms and frequency polygons.
- A frequency distribution consists of a series of pre-determined values or intervals and the frequency with which these values occur.
- A symmetric (non-symmetric) distribution is one in which the frequencies are equal or almost equal 80.95 (unequal) on either side of a center point.
- Relative frequency is the proportion or percent of the total number of observations falling into each interval in a frequency table.
- A histogram is a graphical representation of either a frequency or a relative frequency table.

1.6 References/Further Readings

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

Weiss, a.l. (1996) *Elementary Statistics,* 3rd edition. Addison- Wesley Publishing Co. Inc.

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Jefferies, P.M.(1995) *Mathematics in Nursing,* 4th edition, Bailliere Tindall, London: Cassell and Collier, Macmillan Publishers Ltd.

UNIT 2 CUMULATIVE FREQUENCY

Unit Structure

2.0 Introduction

This unit concerns another significant and fundamental way in which raw data are displayed. This involves summing the frequencies of all classes representing values of data less than the specified class limit.

You will learn how this concept is displayed in a table and as a graph. You will also learn how an associated concept (relative cumulative frequency) is obtained as well as the interpretation of both of them.

2.1 Intended Learning Outcomes (ILOs)

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

- Define the following terms:

a. Cumulative frequency; and
b. Relative cumulative frequency

- Construct cumulative frequency and relative cumulative frequency tables for a set of raw data.
- Interpret a cumulative frequency curve (or ogive).

2.2 Main Content

2.2.1 Class Boundaries

You will observe that this concept was discussed and used in the previous unit. It is a method of showing the classes on the horizontal axis of a histogram. Recall that the lower-class boundary of a class is the number halfway between the lower-class limit of the class and the upper-class limit of the next higher class.

For instance, consider the classes 40-49, 50-59 and 59-60 of the ages of 40 cancer patients given below (see Table 2.1). We have that

Ages of patients	***Frequency (No. of patients)***	***Relative frequency***
30-39	3	0.075
40-49	1	0.025
50-59	8	0.200
60-69	10	0.250
70-79	7	0.175
80-89	7	0.175
90-99	4	0.100
	40	1.000

Lower class boundary $= \frac{49+50}{2} = 49.5$

Upper class boundary $= \frac{59+60}{2} = 59.5$

Table 2.1: Frequency and Relative Frequency Distributions for the ages of cancer patients

You will need to apply the class limit or class boundary in plotting cumulative frequency curves.

2.2.2 Cumulative Frequency

A cumulative frequency is obtained by summing the frequencies of all classes representing values less than the specified class limit. Referring to Table 7.1, we can find the cumulative frequency of the number of cancer patients whose ages are less than 50 years. You will see that the cumulative frequency is $3 + 1 = 4$. This means that *four* of the patients are younger than 50 years.

You will also notice that the number of cancer patients whose ages are less than 80 is given by the cumulative frequency $3 + 1 + 8 + 10 + 7 = 29$. In the next part of the unit, we will discuss a related concept.

2.2.3 Cumulative Relative Frequency

We find a *cumulative relative frequency by dividing the corresponding cumulative frequency by the total number of pieces of data.* You will notice from Table 7.1 that the cumulative relative frequency of the number of cancer patients whose ages are less than 50 is obtained as follows:
Cumulative relative frequency is $4/40 = 0.10$
You may interpret this information by saying that 10% of the cancer patients are less than 50 years.

2.2.4 Ogive

A graph of the cumulative frequency distribution for any collection of data is called *cumulative frequency curve or ogive*
Using Tables 2.1 and 2.2, we can construct an ogive for the ages of the cancer patient's data. You need to be aware that a point is plotted above each class limit at a height equal to the cumulative frequency or cumulative relative frequency. Then you will join all the points with connecting lines. The ogive you will obtain is displayed in the figure below.

Table 2.2: Cumulative frequency and cumulative distribution for the ages of cancer patients.

Ages of Patients	***Frequency (No. of patients)***	***Cumulative Frequency***	***Relative***	***Cumulative Relative frequency***
30-39	3		0.075	
40-49	1	4	0.025	0.10
50-59	8	12	0.200	0.30
60-69	10	22	0.250	0.55
70-79	7	29	0.175	0.725
80-89	7	36	0.175	0.90
90-99	4	40	0.100	1.00

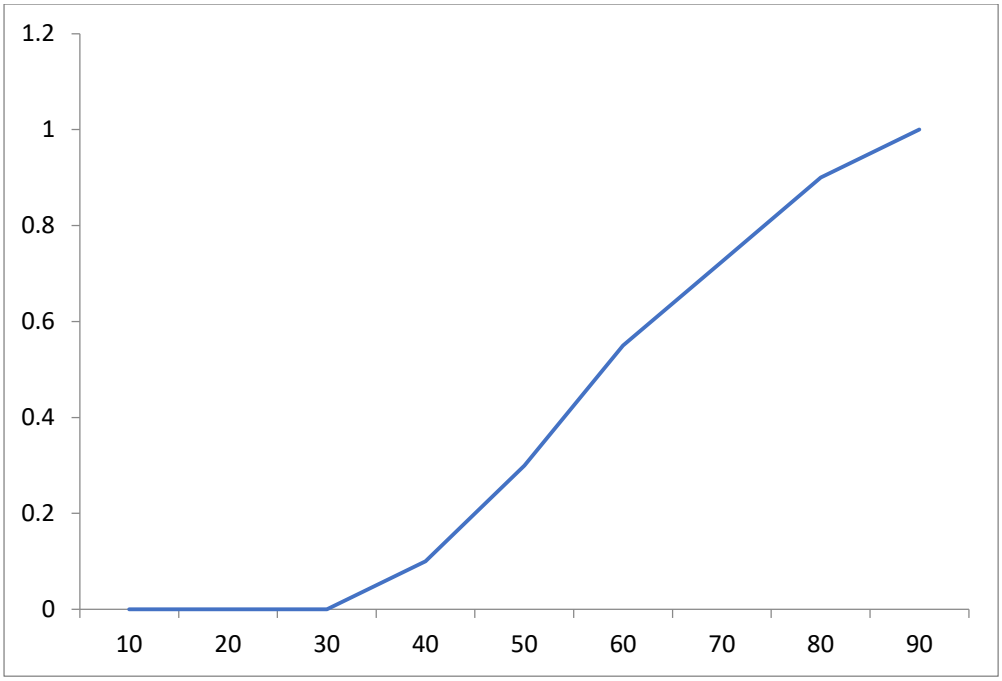


Figure 2.1: Cumulative frequency curve or ogive of ages of cancer patients

2.3 Self-Assessment Exercise(s)

Using the data in Exercise 6.5.1, determine the cumulative frequency and the relative cumulative frequency for the given data.

2.4 Conclusion

In this unit, you have learned the procedure for displaying cumulative frequency or cumulative relative frequency distribution as a graph. You also learned how to interpret a collection of data through this concept.

2.5 Summary

You have learned the procedures for presenting and summarizing raw and unorganized set of data. You also learned how to interpret these data meaningful in terms of graphs.

2.6 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses*, 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics*, 12th edition, E.Arnold, London.

Weiss, A.L (1996) *Elementary Statistics*, 3rd edition, Addison-Wesley Publishing Co. Inc.

UNIT 3 MEASURES OF LOCATION OR CENTRAL TENDENCY

Unit Structure

3.0 Introduction

In two of the preceding units, you have learned that a collection of data can be represented and summarized by graphs, diagrams, frequency and cumulative distributions. You also learned that these techniques are useful in showing some important features of the data.

You will now see in this unit that it is more desirable to describe a collection of data in terms of some other numerical summaries which play crucial roles in the inferential estimation of a population.

In this unit, one of the most important types of numerical summaries will be discussed. These summaries are referred to as *measures of location or central tendency.* They are the mean, median and mode.

3.1 Intended Learning Outcomes (ILOs)

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

- Understand and use the summation operator in computation.
- Define and compute the following measures of central tendency
  a. Mean;
  b. Median; and
  c. Mode.
- Summarize data by means of the measures of central tendency

3.2 Main Content

3.2.1 Measures of Location and Central Tendency

One important descriptive characteristic related to a collection of numerical data is the middlemost value about which another values cluster. These central values are used to locate the center of the frequency distribution and are called *measures of location or central tendency*. They are the mean, median and mode.

We shall discuss their advantages and situations when each one can be used. You will also learn how to define and compute each of the terms. Let us now look at a brief description of one of the most commonly used arithmetic operations in statistics.

3.2.2 Using the Summation Operator

Here we will look at the most commonly used arithmetic operator in statistics. It is used in summing a set of values. The summing operator is denoted by the Greek letter I (sigma). Let us introduce some useful terminologies related to the use of E.

The characteristic or variable of interest in statistics is denoted by $x, y, z,$ etc. A subscript is added to the corresponding symbol denoting a variable of interest; i.e. x_i or y_i or z_i. For instance, if the weight loss in pounds of five patients suffering from tuberculosis is $10, 12, 8.7.5$ then, you can designate these data by x_1, x_2, x_3, x_4, x_5 respectively where x_i refers to the $i - th$ value in the collection of data. Now, if you wish to take the sum of all the weight losses, you write

$$\sum_{i=1}^{n} \mathbf{x_i}$$

The symbol above means that you have taken the sum of the x values starting from i=1 to i=n with n being the number of observations.
If you apply this arithmetic manipulation to the weight losses, you have that the sum of the weight losses is given by

$$52 = 10 + 12 + 8 + 7 + 5 = x_1 + x_2 + x_3 + x_4 + x_5$$
$$= \sum_{i=1}^{5} x_1$$

We will now look at the following examples which illustrate the common uses of the summation operation:

Example 3.1

The ratings of the quality of nursing care in the hospital unit (scale of rating is from 1to 10) by six patients are 3, 2, 4, 1, 6, 7. Using these data, compute the following using $\sum$ notation.

(i) $\sum x_i =$
(ii) $\sum x_i^2 =$
(iii) $\left(\frac{\sum x_i}{4}\right)^2 =$
(iv) $\sum(x_i - 2) =$

Answers to Example 3.1

(i) $\sum x_i = 3 + 2 + 4 + 1 + 6 + 7 = 23$
(ii) $\sum x_i^2 = 3^2 + 2^2 + 4^2 + 1^2 + 6^2 + 7^2$
$= 9 + 4 + 16 + 1 + 36 + 49 = 115$
(iii) $\left(\frac{\sum x_i}{4}\right)^2 = (3 + 2 + 4 + 1 + 6 + 7)^2$
$= \left(\frac{23}{4}\right)^2 = \frac{529}{16}$
(iv) $\sum(x_i - 2) = (3 - 2) + (2 - 2) + (4 - 2) + (1 - 2) +$
$(6 - 2) + (7 - 2)$
(v) $= 1 + 0 + 2 - 1 + 4 + 5 = 11$

3.2.3 Mean

The mean of a set of data is the most commonly used measure of central tendency. It is the arithmetic average of the collection of data i.e. the sum of the collection of data divided by the amount of data. It is represented by

$$\bar{x} = \frac{\sum x_i}{n}$$

Where $\bar{x}$ read x bar'' denote the mean, $\sum x_i$, is the sum of all the x values and n is the number of data in the collection. An illustration of how to find the mean is given below:

Example 3.2

The number of geriatric patients in five urban hospitals in Nigeria is 25, 27, 28, 30, 35. Find the mean of the data for the five hospitals.

Answer

$\bar{x} = \frac{\sum x_i}{n}$

$Mean, \bar{x} = \frac{25+27+28+30+35}{5}$

$= 29$

The mean of the data is 29 patients.

Example 3.3

For the data in 8.2.2.1, compute $\sum(x_i - \bar{x})$

Answer

$$\sum(x_i - \bar{x}) = (25 - 29) + (27 - 29) + (28 - 29) + (30 - 29) + (35 - 29)$$
$$= -4 - 2 - 1 + 1 + 6 = 0$$

You will notice that if the mean is subtracted from all the sample value, the sum of these differences is zero. *The difference between a sample value and the mean is sold to be a deviation.* We shall discuss the importance of this property of the mean in the next study unit.

Example 3.4

Suppose in Example 8.2.2.1, the number of patients in the fifth urban hospital is now 50. What is its effect on the mean of the data?

Answer

$$\bar{x} = \frac{\sum x_i}{n} = 25 + 27 + 28 + 30 + 50$$
$$\bar{x} = \frac{160}{5} = 32$$

You will observe that because of one large data value, the mean has considerably increased. This illustrates one disadvantage of the mean in that it is affected by extreme values, particularly for small number of observations in a sample.

3.2.4 Median

The median of a collection of data is the middlemost measurement when the data are arranged according to size. If the number of observations is odd, the median is the middle measurement, and if the number of observations is even the median is the mean of the two middle observations.

We shall illustrate this concept with the following examples.

Example 3.5

In a research unit of a hospital, the hospital stay (in days) for seven patients are; 12, 13, 15, 16, 17, 18, 20. Determine the median hospital stay for these patients.

Answer

You will first arrange the data in order of increasing magnitude i.e.

$$12, 13, 15, 16, 17, 18, 20$$

The median hospital stay for these patients is 16 days since there are seven observations.

Example 3.6

If given the values; 8, 5, 7, 6, 9, 5, 5, 9, find the median.

Answer

You will first arrange the data in order of increasing magnitude i.e.

$$5, 5, 5, 6, 7, 8, 9, 9$$

Here the median is the average of the 4^{th} and the 5^{th} value

$$\frac{6+7}{2} = 6.5$$

Since there is an even number of data

3.2.5 Mode

The mode of collection of data is the most frequently occurring value

Example 3.7

Find the mode for the following collection of data:

15,15,15,12,13,14,16,12,13

Answer

The mode is 15 since this is the most frequently occurring score or value in the collection of observations.

3.2.6 Choice of a Measure of Central Tendency

Our choice of a measure central tendency depends on our intention of its use. However, the median is the most preferable when the data have the possibility of extreme values. But as the size of the sample increases, the

mean becomes more useful as a descriptive measure. In addition, and for purposes of statistical analysis and inferences, the mean is most useful because it is more amenable to arithmetic manipulations.

When description of qualitative data is required, the more is more useful. Let us consider types of requests made to a nurse by patients in an orthopedic unit. The model request in such a case is that which occurs most frequently.

You should be aware that the mode is scarcely used as a single measure of the central tendency of a collection of data. It is of benefit to use two or three of the measures of central tendency to describe a sample under consideration.

3.3 Self-Assessment Exercise(s)

An investigation was done on survival times for 15 patients following a new treatment of prostate cancer. The times in months were:

25, 16, 20, 25, 30, 35, 20, 27,
35, 40, 28, 40, 24, 25, 30

3.4 Conclusion

In this unit, you have learned how to define and compute measures of central tendency for a set of data. You have also been able to summarize and describe data qualitatively using these measures. Furthermore, you have been able to subject data under consideration to statistical analysis.

3.5 Summary

You have been able to learn in this unit that the general position of a frequency distribution on some scale is measured by an average. You also learned that there are three averages or numerical summaries in common use. These are

(i) The arithmetic mean;
(ii) The median; and
(iii) The mode

The mean is the sum of the measurement divided by the number of measurements in the set. The median of a set of measurements is the

middlemost measure when the values are arranged In order of increasing magnitude. The mode of a set of measurements, is the most frequently occurring value in the set.

Furthermore, you learned that the mean has the disadvantage that it is affected by extreme value and that the median is more preferable as a measure of central tendency in such a case.

In the following Exercises, determine which measure(s) of central tendency is most suitable to describe the Centre of the distribution of the collection of data.

3.6 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

Weiss, A.L (1996) *Elementary Statistics,* 3rd edition, Addison-Wesley Publishing Co. Inc.

UNIT 4 MEASURE OF DISPERSION

Unit Structure

4.0 Introduction

Recall that in the study if unit on presentation of data, you learned that pictorial representations and frequency distribution can describe a collection of unorganized data. In this unit, you will learn that apart from providing a mental image of the frequency distribution of a set of data, there is a means of calculating a measure that reflect the degree of spread of the observed values above the central point.

The most widely used measures of spread or variability are the *range, the variance and standard deviation.* Other measures are *percentage, percentiles, rates and ratios.* Before discussing these concepts, you need to examine the following objectives of the unit.

4.1 Intended Learning Outcomes (ILOs)

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

- Define and compute the following measures of spread;

a. Range
b. Variance
c. Standard deviation

- Select suitable descriptive measure for summarizing data by means of percent's, percentiles.

4.2 Main Content

4.2.1 Measures of Dispersion

In the previous unit, the only set of descriptive measures or numerical measures discussed are namely; the mean, the median and the mode. You saw that these descriptive measures indicate where the center or most typical values of a set of data lies.

However, it is possible for two data sets to have the same mean, the same median or the same mode and yet be quite different in other respects. For example, let is consider two sets of measurements in (a) and (b) below that are centered around the same mean value but do not have the same frequency distribution. You will observe that the diagrams reflect that a greater number of data values are well-scattered about the mean in (a) than in (b).

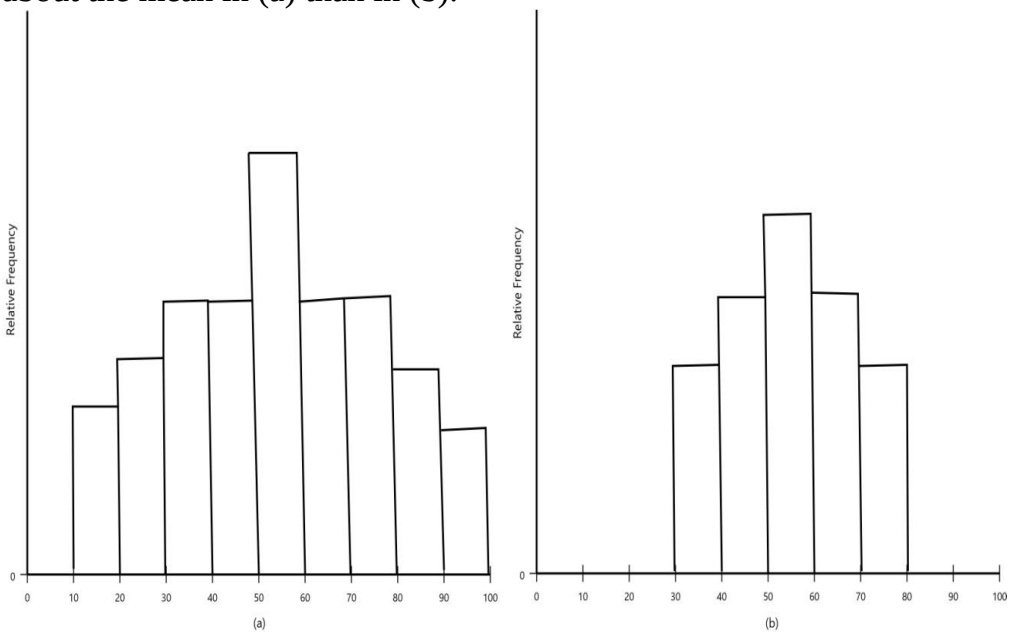


Figure 4.1: *Two histograms with equal mean value but different spread values about the mean.*

We will discuss the most common measures of this spread or scatter in the remaining parts of the unit. These are the range, variance and standard deviation.

4.2.2 Range

For a set of data, the range is the difference between the largest and the smallest scores in the sample. Let us consider the following collection of data: 30 40 50 60 70 80

Here the range is $80 - 30 = 50$
You can observe that the range is very simple to calculate but it does not reflect a correct impression of the spread or variability of the data. Let us consider next the spread of points about the Centre of the following two frequency distributions given in Table 9.2 below:

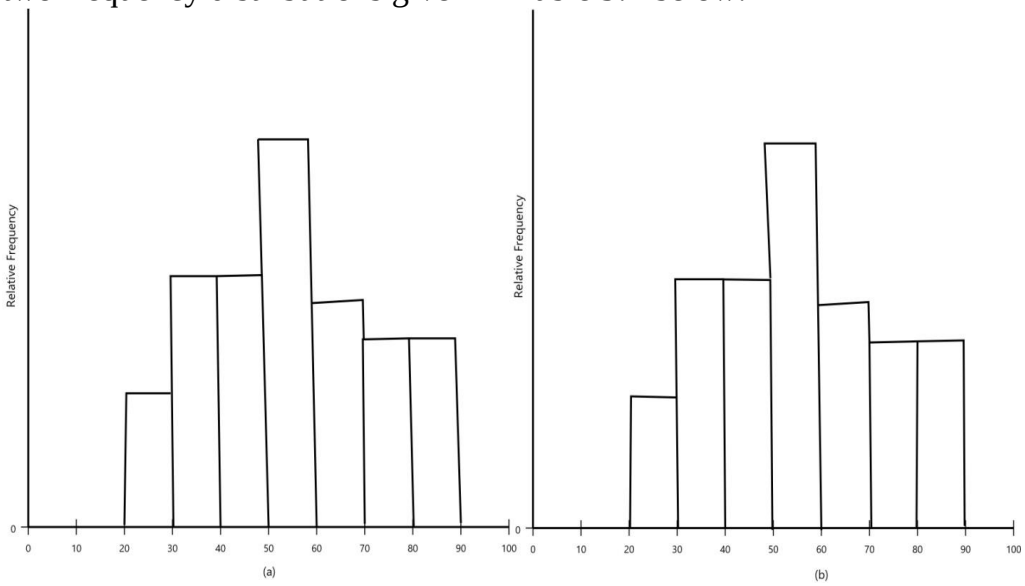


Figure 4.2: *Illustrating Two Histogram of Equal Range*

From the diagrams, you will notice that the range for both (a) and (b) is $90 - 10 = 80$, but there is a remarkable difference between the spread of the points about the centre of the two frequency distributions. The majority of scores clusters about 30 and 80 in (a) while the scores in(b) spread between 20 and 70 more evenly.

You can therefore observe that though the ranges of (a) and (b) are the same yet the scores in (b) have greater spread or variability than in (a).
Another disadvantage of the range as you can observe is that it is affected by extreme scores in collection of data. This means that the presence of one or more extreme scores results in a very large value for the range and this gives a misleading impression of the true spread of the data. Let us now consider the following examples:

Example 4.1

In a study of anxiety of patients towards operation recorded on a scale of score between $0 - 10$, the following data were collected:

$$1, 3, 4, 4, 8, 7, 9, 10, 7, 8$$

Determine the range for the set.
Answer

The largest and the smallest scores are 10 and 1 respectively. Hence the range is $10 - 1 = 9$

Example 4.2

The following set of data, are ages (in years) of 10 recipients of nursing scholarships:

$$25, 21, 22, 20, 19, 30, 27, 28, 32, 18$$

Find the range of data values for the above set.

Answer

The largest and the smallest scores are 32 and 18 respectively. Hence the range is $32 - 18 = 14$.

4.2.3 Variance, Standard Deviation

You will notice from Examples (4.1) and (4.2) that the range provides a misleading idea of the true spread or variability of a collection of data. This is because only two numbers are use in calculating the range. You need to be aware that by making use of all the measurements in a set as well as their deviations from the centre point of a distribution, a more valid measure of variability of the measurements is obtained.
We define a deviation as the *distance between a measurement in the set and the mean value for the set.* Let us look at the illustration of this term using the data in Example 9.2.1.2. Deviations from the mean of the data set are given in the third column of Table 9.1 below.

Table 4.1: *Ages of Recipients of Nursing Scholarships: Deviation from the Mean*

Ages in Years X	*Deviation* $(X_i - \bar{X})$	$(X_i - \bar{X})^2$
18	$(18 - 24.2) = -6.2$	38.44
19	$(19 - 24.2) = -5.2$	27.04
20	$(20 - 24.2) = -4.2$	17.64
21	$(21 - 24.2) = -3.2$	10.24
22	$(22 - 24.2) = -2.2$	4.84
25	$(25 - 24.2) = 0.8$	0.64
27	$(27 - 24.2) = 2.8$	7.84
28	$(28 - 24.2) = 3.8$	14.44
30	$(30 - 24.2) = 5.8$	33.64
32	$(32 - 24.2) = 7.8$	60.84
$\sum x_i = 242$		
$\bar{X} = \frac{\sum x_i}{N} = \frac{242}{10}$	$\sum (X_i - \bar{X}) = 0$	$\sum (X_i - \bar{X})^2 = 215.60$

From Table 4.1, you will observe that the deviations of the individual measurements from the mean give an indication of how spread out the measurements are. You should be aware that the larger the deviations, the more dispersed the measurements are from the Centre of the distribution (the mean)

You are now in a position to determine a measure of variability that takes into consideration the size of all the deviations. You might to simply average the ten deviations and deduce that the set of data having a large average deviation would be more variable than one having a small average deviation. This attempt does not work since the sum of the average deviation is always zero. You will see that the most suitable solution is to square all the deviations and then use the square root of what is obtained to request the 'average' deviation. This is called the standard deviation of the measurements.

We denote the average squared deviation by

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

Where S^2 is called the variance of a set of measurements. You should observe that $(n-1)$ is used instead of n in the formula for reasonably large measurements there is little difference between n and $n-1$.

You should note that the widely used measure of variability is the standard deviation (s) while the variance S^2 is also a very significant indicator of measure of spread of data values. We now illustrate these concepts with the example of the data in table 9.1

Here the square of the deviations of the (variance (s)) ages of recipients of nursing scholarship from the mean is 215.60.

The standard deviation, $S = \ddot{0}\,\overline{s^2} = \sqrt{\frac{S(x_i-\bar{x})^2}{n-1}}$

$$i.e\ = \sqrt{\frac{215.60}{10-1}}\ \rightarrow \ddot{0}\,\overline{23.96} = 4.4895$$

You have now given a suitable description of the location of the center (the mean) of a collection of data as well as how the measurements are spread about the center (the standard deviation).

We now go on to illustrate this concept with another set of examples:

Example 4.3

In a study of the efficacy of two drugs administered on tuberculosis patients, the healing times (in days) was recorded for a number of patients whose ages were between 25-50 years. The mean and standard deviations were reported as follows:

	Drug 1	*Drug 2*
Mean $(\bar{x})$	100	180
Standard Deviation (s)	50	70

Which of the drug would you say is the most efficacious based on the above information?

Answer

Drug 2 has the lower mean of number patients and it has a larger spread of values above the mean. This implies that for this collection of data they are some very short healing times as well as very long ones.

On the other hand, drug 1 has most of its values clustered more closely about the mean 100, showing that the healing times do not change appreciably in either direction from the mean (100). It therefore follows that base on the absence of very long healing times, Drug 1 is adjudged the most efficacious.

Example 4.4

Determine the mean and the standard deviation in study done on survival times for 10m patients following a treatment of cancer. The survival times (in months) are:
24, ,8, 12, 3, 20, 18, 24, 19, 27, 25

Answer: you construct the table below,

Table 4.2: Survival Times: Deviation from the Mean

Survival $(in\ months)$	*Deviation* $(X_i - \bar{X})$	$(X_i - \bar{X})^2$
3	$3 - 18 = -15$	225
8	$8 - 18 = -10$	100
12	$12 - 18 = -6$	36
18	$18 - 18 = 0$	0
19	$19 - 18 = 1$	1
20	$20 - 18 = 2$	4
24	$24 - 18 = 6$	36

24 25 27 $\sum x_i = 180$	$24 - 18 = 6$ $25 - 18 = 7$ $27 - 18 = 9$	36 49 81
$\bar{X} = \frac{\sum x_i}{N} = 18$	$\sum (X_i - \bar{X}) = 0$	$\sum (X_i - \bar{X})^2 = 568$

The mean $(\bar{X}) = 18$
The standard deviation

$$S = \sqrt{\frac{568}{10}} = \sqrt{56.8}$$
$$= 7.5366$$

Example 4.5

Six fathers of premature babies were observed during the initial knowledge of the nature of the births of their babies. The observer rated the father's reaction on a scale from 10 = very despondent to 10 = not despondent, the scores were 2, 2, 3, 4, 5, 8.
Find the;
(i) Mean
(ii) Median,
(iii) Mode,
(iv) Range,
(v) Variance, and
(vi) Standard deviation.

Answer

You will construct the following table where the desired concepts are easily observed:

Table 4.3: Attitude of fathers to premature births: determination of mean, median, mode, range, variance and standard deviation.

Score (X_i)	**Frequency**	**Deviations** $(X_i - \bar{X})$	$(X_i - \bar{X})^2$
2	2	$(2 - 4.25) = -2.25$	
3	1	$(3 - 4.25) = -1.25$	$2(5.0625)$
4	1	$(4 - 4.25) = -0.25$	1.5625
5	1	$(5 - 4.25) = 1.75$	0.0625
8	1	$(8 - 4.25) = 3.75$	3.0625
$\sum x_i = 25$	6		14.0625

$\bar{X} = \frac{\sum x_i}{N} = \frac{25}{6} = 4.25$			$\sum (X_i - \bar{X})^2 = 28.875$

(i) The mean is 4.25

(ii) To obtain the median, we arrange the scores in ascending order of magnitude as follows: 2, 2, 3, 4, 5, 8, Since there are an even number of scores in the set (n=6), the median is the average of the third and the fourth values i.e., $\frac{3+4}{2} = 3.5$

(iii) The mode is the most commonly occurring score i.e., 2.

(iv) Since the largest and the smallest scores are 8 and 2 respectively, the range is $8 - 2 = 6$.

(v) The variance is

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{28.8750}{5} = 5.7770.$$

(vi) The standard deviation

$$s^2 = \sqrt{s^2} = \sqrt{(X_i - \bar{X})^2} = \sqrt{5.7770} = 2.4083$$

4.2.2 Percentiles

In the previous part of the study unit, you have seen how data may be summarized by tables, graphs and by means of numerical summaries like the mean and the standard deviation. You will now see how data may be summarized by another measure by the percentiles. This measure involves the representation of data in relative form. This measure involves the representation called the percentile facilitates the comparison between two sets of data.

A percentile shows the relative position of any individual score with respect to all scores in the collection. As an example, let us compare the performance of nurse X with those of other nurses in a qualifying examination. If you wish to determine X's score in terms of percentiles, the number of scores below X's score is first determined.

You then divide this number by the total number of scores in the collection and multiply by 100 to convert to percent i.e.

$$\text{Percentile} = \frac{Number\ of\ scores\ less\ than\ given\ score}{Total\ number\ of\ scores} \times 100$$

If 80 nurses took the qualifying examination and 16 nurses scored lower than nurse *X*, percentile score of *X* is given by

$$\frac{16}{80} \times 100 = 20\%$$

On the other hand, if 64 nurses had scores lower than $X's, X's$ percentile would be

$$\frac{64}{80} \times 100 = 80\%$$

This means that a percentile score of 80 shows that X is in the top 20% of the group.
By this example you should be aware that the percentile is a numerical means which facilitates the comparison of two scores or sets of scores.

4.3 Self-Assessment Exercise(s)

An investigation was conducted in comparing two techniques for measuring blood pressure. Twenty patients underwent both method X and method Y. for the groups, the mean blood pressures were equal while the standard deviation for the measurements using Y was twice as large as for those using method X. state the more efficient method and the reasons for your choice.

4.4 Conclusion

In this unit, you saw that certain numerical values are necessary as descriptions of the frequency distribution of a collection of data.
You learned that the most important of these are usually the mean and the standard deviation. In addition, you learned that the mean alone is rarely used. Rather, apart from taking into consideration the center of a frequency distribution, the measure of spread (or variability) it displays about the center is significant. In essence, you should be aware of not only the average of a collection of data but the spread of data around it.

4.5 Summary

The critical concept you learned in this unit are:

- Numerical measures are used to describe a set of data.
- Numerical summaries locate the center of distribution of a collection of data and the spread of data about this location.
- The most common measure of variability are the range, variance and standard deviation.
- A percentile shows the relative position of an individual score with respect to all measurements in the collection.

4.6 References/Further Readings

Jefferies, P.M.(1995) *Mathematics in Nursing,* 4th edition, Bailliere Tindall, London: Cassell and Collier, Macmillan Publishers Ltd.

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

UNIT 5 CORRELATION

Unit structure

5.0 Introduction

A problem frequently faced in statistics is how to describe a relationship between two or more variables. For instance, is there a relationship between scores in a nursing qualifying examination and score in a general examination. You need to determine if and how these variables are related.

Some widely used methods for examining the relationship between two or more variables and for making predictions are correlation analysis and regression analysis.

In this unit, we will discuss correlation in terms of its meaning, interpretation and limitation. We will also discuss some of the significant correlation procedures.

5.1 Intended Learning Outcomes (ILOs)

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

- Recognize linear and curvilinear relationships

- Display a set of data by means of scatter diagram
- Define correlation
- Discuss the interpretation of correlation
- Discuss the limitation of correlation
- Apply the most commonly used correlation procedures

5.2 Main Content

5.2.1 Correlation: Meaning and Interpretation

You may have noticed from the introduction that there are several correlation procedures available. They provide the same type of information on the direction and the magnitude of the relationship between variables.

You should be aware that several correlation procedures are needed because different investigations involve different types of variables with the use of different measuring scales. In addition, you need to be aware that despite the number of different techniques, the same meaning and interpretation are obtained.

We will now give a general discussion on the meaning, interpretation and limitations of correlation.

5.2.2 Data Arrangement

In a correlation study, data are usually arranged in pairs (X_i, Y_i). For example, let us see how to represent the following information as data layout for correlation study.

A test considered the measurement of the aptitude of some applicants for admission to a nursing school.

Estimation of the reliability over time of the process involved 10 applicants and was given twice with two-week period. The scores for the applicants are now set out in the following layout:

Table 5.1: *Data layout for Correlation Study*

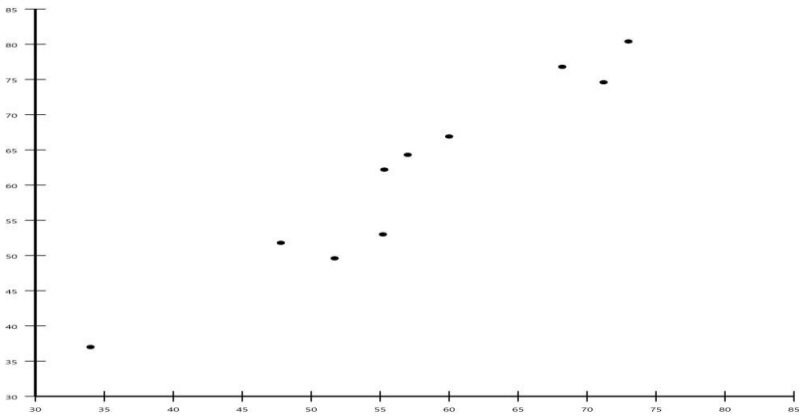
Applicant	***Scores Week 1***	***Scores Week 2***
1	80	75
2	40	50
3	75	70
4	50	50
5	30	35
6	50	45
7	75	80
8	60	65
9	65	65
10	80	85

5.2.3 The Scatter Diagram

The initial step in the investigation of a relationship between variables is a graphical display the data. This display is referred to as a scatter diagram and this gives you a visual image of the relationship studied.

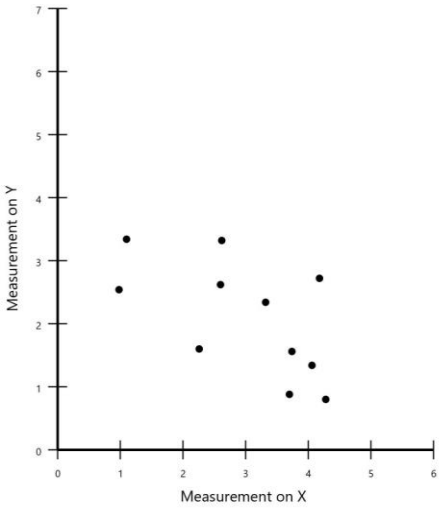
You plot each of the n pairs of points (X, Y) on the graph with the X's and Y's being plotted on the horizontal and vertical axes respectively. You should try to obtain the scatter diagram showing the above point measurements in Table 10.1 as given in Figure 10.1 below

Values of X



corresponding to large or small values of Y. you need to be aware that there is a positive linear relationship between X and Y.

You should also be aware that in a scatter diagram the relationship between X and Y could be negative linear or curvilinear as shown respectively in figures 10.2, 10.3 and 10.4 below. It is possible that you obtain the case where no relationship exists between the variables X and Y.



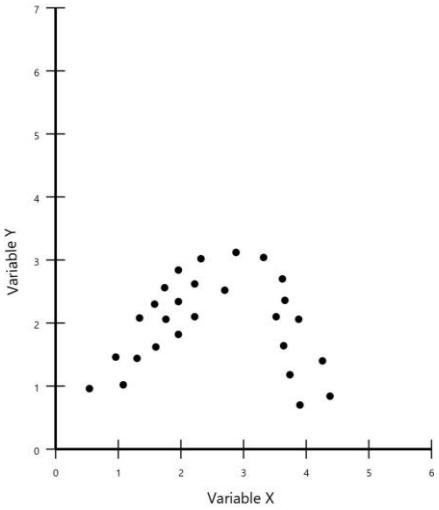


Figure 5.2: A negative linear relationship between *X* and *Y*

Figure 5.3: A curvilinear relationship between *X* and *Y*

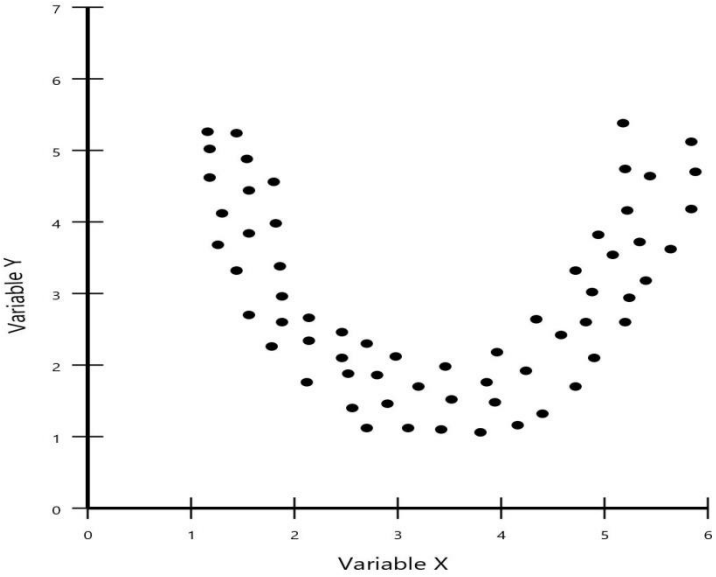


Figure 5.3: A curvilinear relationship between X and Y

5.2.4 Numerical Representation of Relationships between Variables

Apart from the study of relationships through the scatter diagrams, you need to be aware that a numerical representation of such relationship exists. This is called *correlation coefficient which is the magnitude or strength of the relationship between variables.*

Some of the most widely used correlation coefficients include the *Pearson's Product Moment Correlation Coefficient (r) and Spearman's Rank Order Correlation Coefficient (rs).* We will discuss these procedures in the next part of this study unit.

5.2.5 Pearson's Product Moment Correlation Coefficient (r)

This coefficient which measures the linear relationship between two continuous variables, is an index number with values ranging from -1 to +1

You should be aware that a correlation of value 0 is an indicator that no relationship exists between the two variables. In addition, you should note that correlation lie usually in the range -1 to +1 with perfect correlation rarely obtained.

Product Moment Correlation Coefficient is computed using the following formula:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Where,

$$S_{yy} = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$S_{xx} = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$S_{xy} = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

Let us illustrate this computation using the following example;

Example 5.1

A nurse tutor investigated the degree of relationship between students' scores on a battery of personality tests and performance in nursing school and gave a random sample of applicants for nursing school a personality inventory evaluation. The scores on the battery of tests range from 0 to 10. Grade point average in the school was recorded for each student. The data are shown below:

Student	***Personality score***	***GPA***
1	6.0	2.0
2	8.0	3.5
3	7.0	3.0
4	5.0	1.5
5	4.0	1.0
6	3.0	0.5

7	2.0 $\sum X = 35.0$ $\sum X^2 = 203.0$ $\sum XY = 75$	0.5 $\sum Y = 12.0$ $\sum y^2 = 29.0$

Answer

To compute correlation coefficient, we calculate S_{xx}, S_{yy}, S_{xy}

$$S_{xx} = \sum X^2 - \frac{(\sum X)^2}{n} = 203 - \frac{1225}{7}$$
$$= 203 - 175 = 28.0$$
$$S_{yy} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 29 - \frac{144}{7}$$
$$= 29 - 20.6 = 8.4$$
$$S_{xy} = \sum XY - \frac{(\sum X)(\sum Y)}{n} = 75 - \frac{420}{7}$$
$$= 75 - 60 = 15$$
$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{15}{\sqrt{28.0 \times 8.4}}$$
$$r = 0.98$$

5.2.5 Interpretation of Correlation Coefficient

One simple method of interpreting correlation is by squaring the correlation coefficient which indicates the percentage of variation in one variable that is deducted from the variation of the other variable.

In Example 10.2.4.1, the square of the correlation is $(0.88)^2 = 0.7744$. This implies that 77.44% of the variability in the personality scores is attributable to variable in the student's GPA. Another interpretation is that, given a student's GPA, 77.44% information is available in predicting his personality score.

You however need to be aware of the following drawbacks of the correlation coefficient:

(i) A correlation must be greater than 0.7 before 70% of the variation in one variable may be attributed to a variation in the other variable.
(ii) Extremely high correlation (e.g., 0.95) suggest the likelihood of one quantity being correlated with a quantity of which the first is a component.

There is the need to accompany the computation of the coefficient of correlation by a level of significance. However, this aspect is delayed until we discuss simple significance tests in unit 17.

5.2.6 Precautions in use in the Interpretation of Correlation

You need to be aware of the most important rules of the interpretation of correlation. These are:

1. Correlation does not mean causation i.e.; correlation studies do not prove that one variables causes another. For example, X correlated with Y implies Y correlated with X. so you cannot state that X causes Y, neither can you say that Y causes X.
2. Correlation only applies to the range of values observed for the two variables.
3. Interpretation of correlation is dependent on the particular investigation and the judgment of the investigator and the consumer.

5.2.7 Correlation Ratio (h)

Another measure of the relationship between continuous variables is the correlation ratio denoted by *h.* it is used when the relationship between two continuous variables is curvilinear. Its value ranges from -1 to +1 with its magnitude being an indication of the degree of association between the two variables. You also need to accompany this relationship by means of a scatter diagram.

5.2.8 Spearman's Rank Order Correlation Coefficient

The Spearman's Rank Order Correlation Coefficient is a non-parametric or distribution-free statistical technique. It is used when the assumptions underlying the classical techniques fail. It is convenient to use this correlation coefficient when data values are assigned ranks.

Let us consider the following examples where the spearman's rank order correlation coefficient is used:

In a study involving the quality of clinical performance for six student nurses in which evaluation was conducted by two different observers, the ranking of student performance was from 1 to 6 by each observer.

Nurse No	***Rank***	
	Observer 1	***Observer 2***
1	2	5
2	3	1
3	4	6
4	5	2
5	1	3
6	6	4

You should note that the Spearman's Rank Order Correlation coefficient is used to measure the degree of relationship between the scoring of the two observers.

The computation formula for the Spearman's Rank Order Correlation Coefficient is given by

$$r_x = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

With *d* being the difference between ranks for each individual and *n* is the number of pairs of ranks. To compute this coefficient for the previous example we have that

$$\sum d^2 = (-3)^2 + (2)^2 + (-2)^2 + 3^2 + (-2)^2 + 2^2$$
$$= 9 + 4 + 4 + 9 + 4 + 4$$
$$= 34$$
$$r_x = 1 - \frac{6\sum d^2}{n(n^2-1)}$$
$$= 1 - \frac{6(34)}{6(36-1)} = 1 - \frac{204}{210}$$
$$= 1 - 0.97 = 0.03$$

The interpretation of the Spearman's Rank Order Correlation Coefficient is that it takes values from *-1 to +1* with zero indicating no linear relationships between the ranks. It is also usual that this coefficient is subjected to some hypothesis testing.

5.3 Self-Assessment Exercise(s)

The following data represent systolic blood pressures readings (in mmHg) on six patients read by two nurses using the same instrument.

Patient	*Nurse 1 (X)*	*Nurse 2 (Y)*
1	130	125
2	140	135
3	136	135
4	154	160
5	120	160
6	165	125

Display the data graphically on a scatter diagram and determine the magnitude of the association between the two sets of scores using n.

5.4 Conclusion

In this unit, you saw that the correlation coefficient is a useful measure of the degree of association between two (or more) variables but this is valid only when a straight line adequately describes this relationship.
You also learned that the error of this estimation may be large even when the correlation is high. You also saw that evidence of association is not necessarily that of causation and that influence of other factors needs be taken into consideration so as to significantly interpret correlation coefficients.

5.5 Summary

In this unit, the following concepts were presented:

(i) In a correlation study, the initial step is the display of a scatter diagram.
(ii) A scatter diagram involves plotting each pair of observations (X_i, Y_i) as a single point on a graph.
(iii) If the points seem to fall almost on a straight line and the X and Y values increase or decrease correspondingly the relationship is positive linear. If X values increase as Y values decrease, the relationship is negative linear.
(iv) If the points appear to lie on a curve, the relationship is curvilinear
(v) If as X values increase or decrease as Y values is unchanged no relationship exists.
(vi) Correlation coefficient measures the magnitude of the relationship between X and Y.
(vii) Correlation coefficients are index numbers whose values range from -1 to +1. Zero indicates no relationship between variables.
(viii) Perfect correlations rarely occur in practice.

(ix) The square of the correlation coefficient measures what knowledge of the first variable gives information about the second. It is the percentage of variation in one variable, which is explained by the variation of the other.
(x) Correlation does not imply causation.
(xi) Pearson's Product Moment Correlation Coefficient (r) measures the linear relationship between two continuous variables.
(xii) Spearman's Rank Order Correlation Coefficient measures the degree of association between two variables that have been ranked.

5.6 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

Weiss, A.L (1996) *Elementary Statistics,* 3rd edition, Addison-Wesley Publishing Co. Inc.

MODULE 3 PROBABILITY DISTRIBUTION

Introduction

A probability distribution is an idealized frequency distribution. A frequency distribution describes a specific sample or dataset. It's the number of times each possible value of a variable occurs in the dataset.

The number of times a value occurs in a sample is determined by its probability of occurrence. In probability theory and statistics, a probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events.

UNIT 1 REGRESSION

Unit Structure

1.1 Introduction

In the previous study unit, you learned one of the most commonly used method for examining the relationship between two or more variables.

This unit is concerned with a topic closely related to the one discussed in the previous unit. Here you will learn the statistical technique of or predicting the value of one variable given the value of a second variable. This technique is referred to as regression analysis.

In regression analysis, the relationship between two variables (or more) entails fitting a line or a curve to pairs of data points. You will therefore need to review some mathematical concepts involving a straight line. However, before proceeding in that direction let us examine what you will learn in this unit as stated in the objectives hereunder:

1.2 Intended Learning Outcomes (ILOs)

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

- Describe in words the use of regression analysis
- Discuss precautions in the use of regression
- Carry out a simple linear regression analysis

1.3 Main Content

3.1 Linear Equation with One Independent Variable

You will need to review linear equations with one independent variable in order to understand linear regression.

The first step in this direction is for you to observe that the general form of a linear equation with one independent variable is given by

$$b + b_i x$$

Where b and b_i are constants (or fixed numbers), x and y are respectively the independent and dependent variables.

The next step is to see that when the graph of a linear equation is displayed, you will obtain a straight line. We draw the graphs of the following three linear equations (see Figure 11.1 below) to illustrate this concept:

$$y = x + 4$$
$$y = -2x + 3$$
$$y = 3x - 1$$

You need to obtain table of values for each equation as follows:

$$y = x + 4$$

x	-4	-3	-2	-1	0	1	2	3	4
y	0	1	2	3	4	5	6	7	8

$$y = -2x + 3$$

x	-4	-3	-2	-1	0	1	2	3	4
Y	11	9	7	5	3	1	-1	-3	-5

$$y = 3x - 1$$

X	-4	-3	-2	-1	0	1	2	3	4
Y	-13	-10	-7	-4	-1	2	5	8	11

Then plot the points given in the tables and join the points with smooth straight lines.

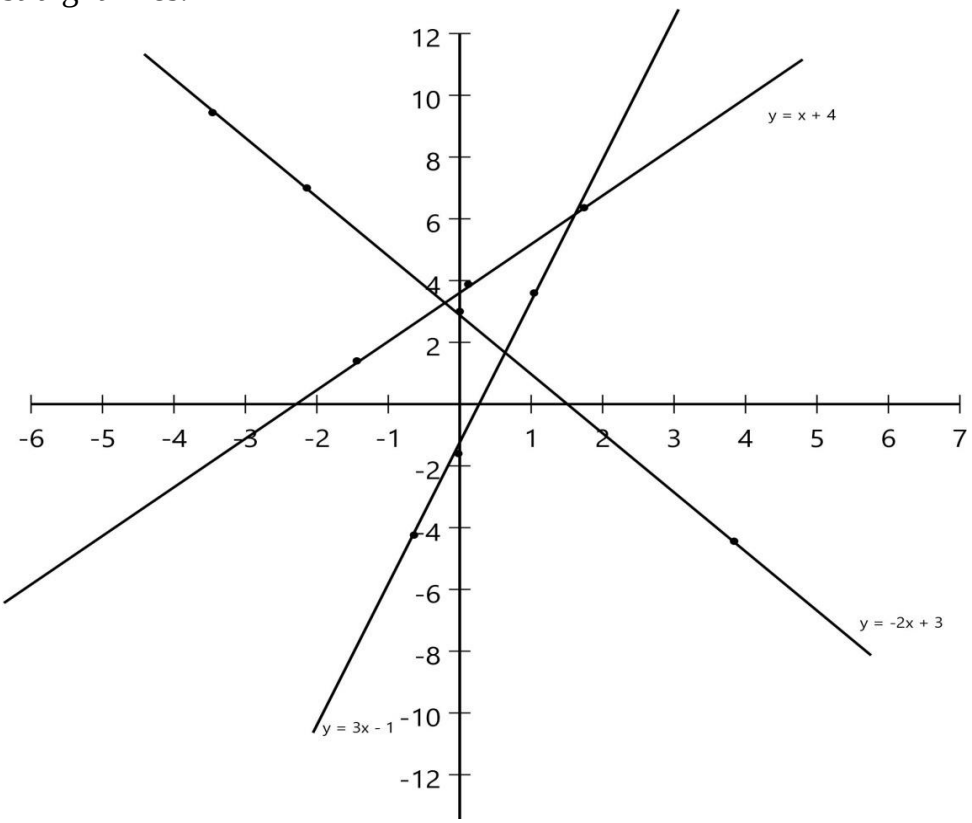


Figure 1.1: Straight line graphs of three linear equation

3.2 Intercept and Slope

For a linear equation $y = b_0 + b_1x$, the numbers b_0 and b_1 have geometric interpretation. b_0 is the y-value at which the straight-line graph of the equation cuts the y-axis. This number b_0 is said to be the *intercept* of the graph on the y-axis.
b_1 indicates how much the y-value on the straight-line increases (or decreases) when the x-value increases by *1* unit. b_1 is called the *slope* of the graph. We illustrate these terms using the following examples:

Example 1.1

Given the equation $y = 25 + 20x$
(a) Find the y intercept and slope of the equation.

(b) Interpret the y-intercept and slope in terms of the graph of the equation.

Answer

(a) The y-intercept $b_0 = 25$ and the slope $b_1 = 20$

(b) The y-intercept $b_0 = 25$ is the y-value at which the straight line $y = 25 + 20x$ cuts the x-axis.

The slope $b_1 = 20$ indicates that the y-value increases by 20 units for every increase in x of *1* unit.

$$y = 25 + 20x$$

x	0	1	2	3	4	5
y	25	45	65	85	105	125

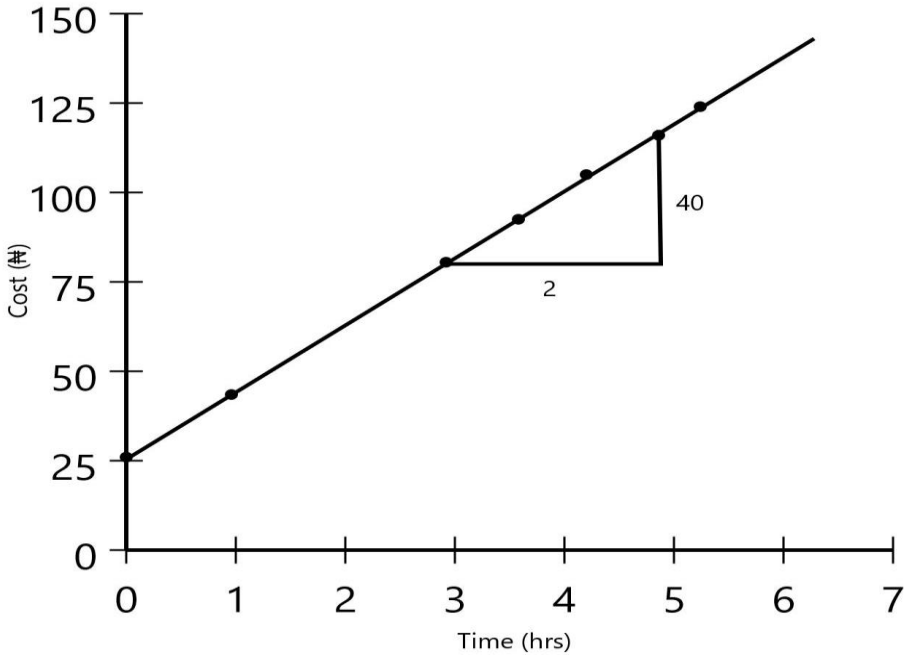
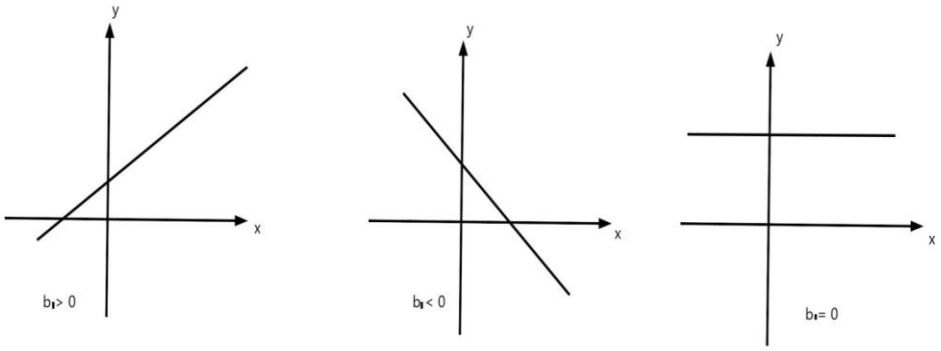


Figure 1.2: *Intercept and Slope of a Graph*

3.3 Graphical Interpretation of Slope

You need to know that a straight line is determined by any two distinct points that lie on the line. An implication of this is that you need to substitute two different x-values into the equation to get two distinct points and then you connect those two points with a straight line.

Another point you need to notice is that the straight-line graph of the linear equation
$y = b_0 + b_1x$ slopes upward if $b_1 > 0$, slopes downward if $b_1 < 0$ and is horizontal if $b_1 = 0$ as shown in Figure 11.3 below.



3.4 Regression Equation

Regression analysis usually starts with a plot of the data on a scatter diagram. You then observe if some of the data points fall almost on a straight line or not. This indicates that the relationship could be linear or not. The next step is to determine the equation of the line.

Since you could draw many straight lines though the cluster of data points, you need a method to choose the best-fitting line. The statistical procedure for finding this line of best fit is called the *method of least squares* and the line so obtained is called the *regression line.* The equation of the line is referred to as *regression equation*. The formal procedure for deriving the method of least squares is beyond the scope of this course.

However, we will employ the result in the following problem to illustrate the principles involved.

$$S_{xy} = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

$$S_{yy} = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$S_{xx} = \sum X^2 - \frac{(\sum X)^2}{n}$$

Example 1.2

A study was conducted so as to weights of premature infants based on their ages in weeks. The following observations were collected on 10 infants.

Infant	***Age (in weeks)***	***Weight (in pounds)***
1	6	6.0
2	3	2.5
3	2	2.0
4	2	2.0
5	1	2.0
6	3	3.0
7	4	4.0
8	5	4.5
9	7	6.5
10	4	3.5
$\sum X^2 = 169.0$	$\sum X = 37$	$\sum Y = 36.0$
	$\sum XY = 160.5$	$\sum Y^2 = 144.0$

Determine the equation of the line relating weights of premature infants to age (in weeks).

Answer

The simple regression line is given by

$$\frac{\dot{U}}{y} = \frac{\dot{U}}{b_o} + \frac{\dot{U}}{b_1}\bar{x}$$

With the least square estimate being

$$\frac{\dot{U}}{b_1} = \frac{S_{yy}}{S_{xx}} : \frac{\dot{U}}{b_o} = \bar{y} - \frac{\dot{U}}{b_1}\bar{x}$$

From the data above we compute

$$\bar{y} = \frac{SY}{n} = \frac{36}{10} = 3.6$$

$$\bar{x} = \frac{SX}{n} = \frac{37}{10} = 3.7$$

$$S_{yy} = SY^2 - \frac{(SY)^2}{n} = 144 - \frac{36^2}{10} = 144 - 129.6 = 14.4$$

$$S_{xx} = SX^2 - \frac{(SX)^2}{n} = 169 - \frac{37^2}{10} = 144 - 136.9 = 32.1$$

$$S_{xy} = SXY - \frac{(SX)(SY)}{n} = 160.5 - \frac{133.2}{10} = 27.3$$

We now have

$$\frac{\dot{U}}{b_1} = \frac{27.3}{32.1} = 0.85$$

$$\frac{\dot{U}}{b_o} = 3.6 - \frac{27.3}{32.1}(3.7) = 3.6 - 0.85(3.7) = 3.6 - 3.15$$

$$= 0.450$$

$$\frac{\dot{U}}{y} = 045 + 0.85x$$

You should note from the example above that in the context of regression analysis that in equation

$$\frac{\dot{U}}{y} = \frac{\dot{U}}{b_o} + \frac{\dot{U}}{b_1}\bar{x}$$

y is called the *response variable* while x is referred to as the *predictor or explanatory variable.* This is because x is used to predict or explain the values of the response variable.

3.5 Precautions on the use of Linear Regression

You have seen that the concept behind finding a regression line is based on the assumption that the data points are scattered about a straight line. But in some cases, data points may be scattered about a curve instead of a straight line.

In such cases, techniques are available for fitting curves to data points showing a curved pattern. These techniques involve curvilinear regression.

You need to be aware also that a measure of scatter or variation of the observed points (y) about the regression line may be estimated by the following formula:

$$s_{y,x}^2 = \sum (y_i - \widehat{y_2})^2$$

Where n is the number of pairs (x_t, y), is the observed values of y and p, is the value of y predicted by the regression line.

The positive square root of s^2x is referred to as the *standard error* of the estimate. It is a measure of average deviation of the observed values (y) from the values (y) predicted by the regression line.

In conclusion, there is the need for you to evaluate the sample regression line so as to determine if adequately describes the relationship between the variables X and Y.

You will accomplish this through tests of hypothesis on the true slope of the line. Again, this discussion is delayed until we discuss some basic

elements of inferential statistics in later units. We mention here passing those other types of regression exists i.e., multiple linear regression.

1.4 Self-Assessment Exercise(s)

The admissions panel of a school of nursing wished to formulate an equation for predicting a student clinical performance based on a set of personality-IQ tests given to students on application for admission the school. Scores on the personality-IQ test along with grade point average for clinical performance we obtained for a random sample of 8 nursing students. The data recorded are as follows:

Student	**Personality-IQ Test Score**	**Clinical Grade Point Average**
1	48	**2.6**
2	45	2.4
3	60	3.4
4	55	3.0
5	40	2.1
6	25	1.5
7	38	1.9
8	30	1.7

(i) Plot the scatter diagram

(ii) Determine the equation of the line relating clinical grade point average to scores on the personality-IQ test.

(iii) Test the hypothesis that there is a significant linear relationship between clinical grade point average and scores on the personality-IQ test (use the 1% level of significance).

1.5 Conclusion

In this unit, one other significant and useful measure of the degree of association between two characteristics of a population was discussed. You have learned that this relationship is valid only when it is adequately described by a straight line.

You also learned that the equation to this line is called *the regression equation* and that the equation allows the value of one characteristic to be estimated when the value of another characteristic is known.

You also saw that in many situations particularly in medicine and nursing practice association between the prevalence of disease, mortality and environmental factors are not uncommon. These examples of the

relationships between two or more characteristics of observation lead to procedures of multiple linear or curvilinear regressions.

1.6 Summary

In this unit, the summary of critical concepts learned involve regression analysis which is a statistical technique for predicting the value of one valuable, given the value of a second variable.

1.7 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd Edn., Delmar Publishers Inc. N.Y.

Jefferies, P.M.(1995) *Mathematics in Nursing,* 4th edition, Bailliere Tindall, London: Cassell and Collier, Macmillan Publishers Ltd.

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

UNIT 2 SIMPLE CONCEPTS OF PROBABILITY

Unit Structure

2.1 Introduction

Up to this point and in the previous study units, we have been dwelling in the realm of descriptive statistics. This concerns mainly the techniques of organizing and summarizing data.

In this unit, we shall be building up the fundamentals of inferential statistics. This involves methods of drawing conclusion about a population having regard to the information obtained from a sample of the population.

Since inferential statistics concerns employing information obtained from a part of a population in drawing conclusions about the whole population, there exist a measure of uncertainty in such techniques. It is therefore essential to be conversant with uncertainty in the understanding, development and application of techniques of inferential statistics. The science that deals with uncertainty is probability theory. In this unit, simple concepts of probability theory are discussed. Let us first look at what you will learn in this unit as stated in the objectives hereunder.

2.2 Intended Learning Outcomes (ILOs)

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

- Define terms relating to probability.
- Identify events
- Apply basic rules of probability
- Recognize and explain some simple concepts of probability.

2.3 Main Content

2.3.1 Classical Probability

You should be aware that although most applications of probability theory to statistics inference deal with large populations yet the simple concepts of probability are best illustrated and explained with small populations and games of chance.

You will gradually learn how to apply the basic principles of probability theory to solve problems in inferential statistics in later units with the foundation developed in this unit. We will now give a more lucid definition of probability in the next part of the study unit.

2.3.2 Meaning of Probability

One basic term here is an *event*. You need to understand that an *event is some specified result, which has the likelihood of occurrence or not in an experiment.* For instance, when you toss a coin, the occurrence of a head or tail is likely to occur or not.

You should then note that the *probability of an event is defined as a measure of its likelihood of occurrence.* You now need to note that a probability very close to zero is an indicator of an event unlikely to occur, whereas a probability of 1 (or 100%) indicates that the event is very likely to occur.

2.3.3 Equal-Likelihood Model

Our discussion of classical probability entails the equal-likelihood model. This is the application of the model whereby possible outcomes of an experiment are equally likely to occur.

An illustration of this model goes thus:

Let *N* equally likely outcomes for an experiment be possible. You will then notice that the probability that a specified event occurs is the number of ways that the event can occur, divided by the total number *N* of possible outcomes i.e.

Probability of an event $= \frac{f}{N}$

Let us illustrate this concept with the following example:

Example 1.1

Suppose it is desired to select a very brilliant nursing student from a class of 40 students; the selection being done such that the chosen student is 20 years old. Suppose further that there are only *seven* 20 years old in the whole class.

Answer

The event that student chosen is 20 years old can occur in seven ways because there are only seven students in the class who are 20 years old, hence $f = 7$ for the event. Therefore, the probability that the student is 20 years old is

$$= \frac{f}{N} = \frac{7}{40} = 0.175$$

2.3.4 Probabilities and Percentages

If an experiment consists of choosing one member at random from a finite population, it follows that the probability that a specified event occurs is equal to the relative frequency (percentage) of members of the population that satisfy the conditions prescribed by the event.

2.3.5 Basic Properties of Probability

There are some simple and fundamental properties of probabilities you have to note. These are as follows:

(i) Probability of an event lies between 0 and 1 inclusive.
(ii) Probability of an event not occurring is zero (i.e.an impossible event).
(iii) Probability of an event occurring is 1 (i.e., a certain event).

You need to note that the procedure you have learned for computing probabilities is applicable to experiments with possible outcomes equally likely to occur. If this is not the case, you must use other methods to determine probabilities.

You will learn these other methods later.

Events

You have learned to use the word *event* rather intuitively. The precise definition in probability is that an *event consists of a collection of outcomes*.

For example, a deck of playing cards consists of 52 cards. If you randomly select a card from the pack, exactly one of these 52 cards is obtained. A collection of possible outcomes (i.e., 52 cards) is said to be a *sample space* of the selection.

You will notice that several and distinct events can be attached to this card-selection process. 12.3.1 Notation and Graphical Displays for Events.

You should notice that the conventional method of representing events involves the use of the letters *A, B, C, D* Visual representation of events and relationships between them is done by means of *Venn Diagrams.*
The sample space is shown as a rectangle and the various events are drawn as disks in the rectangle. We have in the diagram below the event *E* in a sample space.

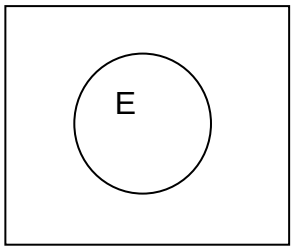


Figure 12.2: *Venn diagram for event E*
2.3.5.1Relationships among Events

You should observe that to each event *E* there exists another event *E'* (complement of *E*) which is the likelihood of *E* not occurring. In addition, you need to be aware that with any two events *A* and *B,* the likelihood of A or B occurring (i.e., A L.) B) exist.
These three new events arising from events, A and B can be illustrated by the following Venn Diagram:

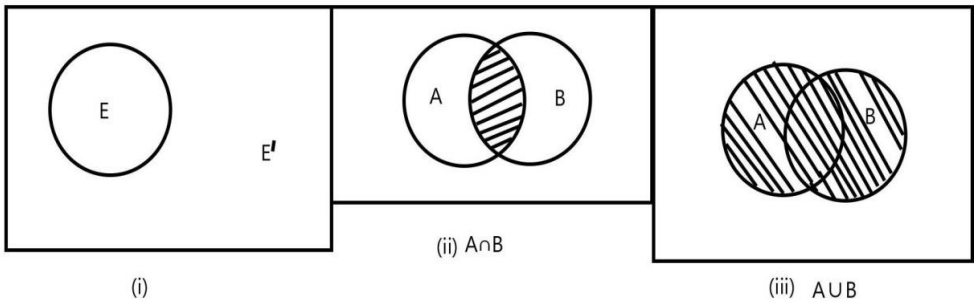


Figure 1.2: ***Venn diagram for E, A ∩ B and A ∪ B.***

2.2.5.2 Mutually Exclusive Events

Two or more events are said to be mutually exclusive if the two of them cannot together occur when an experiment is performed.

We illustrate this concept with the card selection experiment. Suppose
A is the event that the card selected is 8
B is the event that the card selected is 10.

Then, A and B are mutually exclusive since the selection of 8 is not an outcome to that of 10

2.3.6 Some Rules of Probability

In this section, you will learn several rules of probability. Before beginning, you need to be conversant with a significant notation of probability.
Let us consider a balanced die which is thrown once, six equally likely outcomes are possible. If the die comes up even, then this event can occur in three ways i.e if 2, 4, 6 are rolled. You can apply the equal likelihood model to obtain $L = 3 = 0.5$ as the probability of the event that the die comes up even.

$$N6$$

A notation P(A) is used to represent the probability that the event (A) that the die comes up even i.e. you can add P(A)=0.5 as the probability of event A occurring is 0.5
In summary, if E is an event, then P(E) stands for the probability that event E occurs.

2.3.7 Special Addition Rule

If event A and event B are mutually exclusive, we have that

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

In general, if events $A, B, C,$ are mutually exclusive, then

$$P(A\ orB\ or\ C\ or\ \ldots) = P(A) + P(B) + P(C) + \cdots$$

We now illustrate this principle with the following example:

Example 1.2

A compilation of information about five distinct diseases in the general hospital revealed a relative frequency distribution presented in the table below

Disease	***Relative frequency***	***Event***
Cancer	0.087	A
Tuberculosis	0.156	B
Respiratory	0.457	C
Orthopedic	0.184	D
Psychiatric	0.116	E

Table 1.1: Common disease in a general hospital

Answer

As you can observe from Table 12.1, the event of treatment of cancer in this hospital can be represented by A. Events A, B and C are mutually exclusive and so by the special addition rule we have

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$
$$= 0.087 + 0.156 + 0.457 = 0.700$$

You may interpret this rule as follows:
70% of disease treated in the hospital consist of cancer, tuberculosis and respiratory problems.

2.3.7.1 Complementation Rule

The second rule of probability you will learn is the complementation rule which states that the probability that an event occurs equals 1 minus the probability that it does not occur. This may be expressed as the formula

$$P(E) = 1 - P(not\ E)$$

For any event E
You can see the validity of this rule using the Venn diagram below:

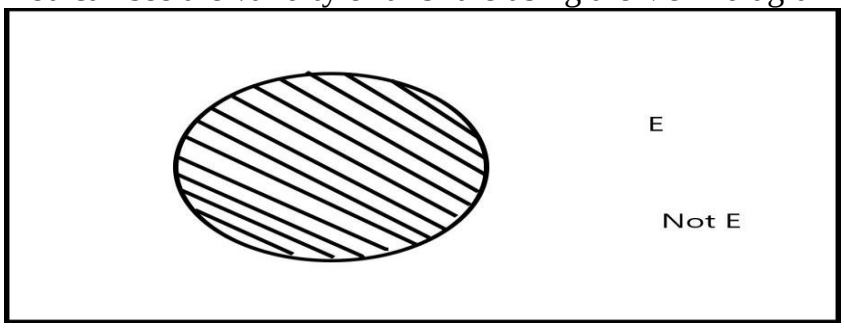


Figure 1.3: An event and its complement.

You find out that the complementation rule is significant in the sense that for an event E, it is sometimes easier to compute the probability that E does not occur than the probability that E does not occur than the probability that it does occur.

2.3.7.2 General Addition Rule

The special addition rule is valid only for events that are mutually exclusive. In the case of events that are not mutually exclusive you will need to use a different rule. This is the general addition rule, which states that:
If *A* and *B* are any two events, then

$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$$

i.e. the probability that either event A or event B occurs equals the probability that event A occurs plus the probability that event B occurs minus the probability that both occur. We will illustrate this rule with the following example:

Example 1.3

Data on admission of patients into a general hospital revealed that in 1999, 82% were male, 21% were under 18 and 10% were males under 18. If a patient admitted in 1999 is randomly selected, find the probability that the patient is either male or under 18.

Answer

Suppose
N=event the patient admitted is male.
E=event the patient admitted is under 18
The event that the patient admitted is either male or under 18 can be represented by (N or E). We wish to find (N and E)

$$\mathrm{P(N)} = 0.82, \mathrm{P(E)} = 0.21$$
$$\mathrm{P(N\ and\ E)} = 0.10$$

By the general addition rule

$$\mathrm{P(N\ or\ E)} = \mathrm{P(N)} + \mathrm{P(E)} - \mathrm{P(N\ and\ E)}$$
$$= 0.82 + 0.21 - 0.10$$
$$= 0.93$$

i.e. 93% of those admitted in 1999 were either male or under 18.

2.3.8 Conditional Probabilities

We now introduce the concept of conditional probability. This is defined as that probability that the event occur under the assumption that another event has occurred.

You may also understand this concept in the following manner. Let *A* and *B* be events, then the probability that *B* occurs given that *A* has occurred is said to be a conditional probability. It is denoted by $\mathrm{P}(\mathrm{B}/A)$. We may illustrate the concept with the following example.

Example 1.4

If a balanced dice is thrown once, six equally likely outcomes are possible. Suppose *K*=event a 3 is thrown, *L*=event the die comes odd. Determine the following probabilities:

(i) P(K), the probability that a 3 is thrown

(ii) P(K/L), the conditional probability that a 3 is rolled given that the die comes up odd

(iii) P(L/not K), the conditional probability that the die comes up odd given that a 3 is not thrown.

$$P(A_1) = \frac{40}{100} = 0.4$$

$$\mathrm{P}(\mathrm{A}_2) = \frac{80}{100} - 0.8$$

$$\mathrm{P}(\mathrm{A}, \mathrm{IA},) = \frac{20}{100} = 0.25$$

In summary, if *A* and *B* are any two events, then

$$\mathrm{P}\ (\mathrm{B}/\mathrm{A}) = \frac{P(A \text{ and } B)}{P(A}$$

This rule states that the *conditional probability that event B occurs given that event A has occurred is equal to the joint probability of events A and B divided by the probability of event A*

2.3.9 Multiplication Rule

In the previous part of this unit, you saw that the conditional probability rule is used for computing conditional probabilities in terms of unconditional probabilities i.e.

$$P(B/A) = \frac{P(A \text{ and } B)}{P(A)}$$

If both sides of this formula is multiplied by P(A), you will obtain a formula for computing joint probabilities in terms of marginal and conditional probabilities i.e.

$$\mathrm{P}(\mathrm{A} \text{ and } \mathrm{B}) = \overline{\mathrm{P}(\mathrm{A})}.\,\mathrm{P}(\mathrm{B}/\mathrm{A})$$

For any two events *A* and *B*

This formula is referred to as the general *multiplication rule.* It states that the probability that both event *A* and event *B* occur equals the probability that event *A* occurs times the probability that event *B* occurs given the event A has occurred.

2.3.10 Statistical Independence

Two events are said to be statistically independent if the occurrence (or non-occurrence) of one of the events does not affect the probability of the other event.

You need to be aware of the following formal definition of the concept:
Event B is said to statistically independent of event A if the occurrence of event A does not affect the probability that event B occurs. Symbolically, we have

$$P(B/A) = P(B)$$

We now illustrate this concept with the following example;

Example 1.5

In the experiment of randomly selecting a card from a deck of 52 playing cards, let
A = event a face card is selected
B = event a King is selected

Determine whether event B is independent of A

Answer
You should observe that the unconditional probability that event B occurs equals

$$P(B) = \frac{F}{N} = \frac{4}{52} = \frac{1}{13}$$

To determine whether event B is independent of event A, you need to compute *P(B/A)* and then compare it with P(B). if P(B/A) = P(B), then event *B* is independent of event *A*.

$$P(B/A) = \frac{F}{N} = \frac{4}{12}$$

Now *P(B/A).P(B)*, so the occurrence of event A affects the probability that *B* occurs. This means that event *B* is not independent of *A*.

2.4 Self-Assessment Exercise(s)

Data on patients admitted for treatment in a hospital in 1991 revealed that 81.3% were male, 16.3% were under 18 and 12.6% were males under 18. If a person admitted in 1991 is selected at random, find the probability that the person is either male or under 18.

Exercise 2.1

A census compiled revealed the following data on the marital status of nursing officers in a particular hospital.

Marital status	**Single**	**Married**	**Widowed**	**Divorced**	**PS.**
Male	M_1	M_2	M_3	M_4	
Female	0.166	0.319	0.012	0.028	0.475
$P(M_i)$	0.093	0.325	0.066	0.041	0.525
	0.209	0.644	0.078	0.069	1.000

If a nursing officer is selected at random

(i) Determine the probability that the nursing officer selected is divorced given that he/she is a male.

(ii) Determine the probability that the nursing officer selected is a male given that he is divorced.

2.5 Conclusion

In this unit, you have learned the simple concepts of probability theory. This is the foundation of future study of inferential statistics.

2.6 Summary

In this unit, the critical concepts that you studied included

- Classical probability which utilizes the equal-likelihood model
- Basic properties of probabilities
- Definition of an events
- Relationships among events
- Mutually exclusive and statistically independent events
- Several rules of probability
- Conditional and unconditional probability.

1. The ages of nursing students in a Teaching Hospital are displayed in the Table below. Suppose one of the students is selected at random, meaning that each student is equally likely to be selected. Find the probability that the student selected is 20 years old.

Table 12:2 Grouped data table for the ages of nursing students in a Teaching Hospital.

Ages (years)	***Frequency***
17	
18	1
19	9
20	7
21	7
22	5
23	3
24	4
25	1
Total	**38**

2.7 References/Further Readings

Jefferies, P.M.(1995) *Mathematics in Nursing,* 4th edition, Bailliere Tindall, London: Cassell and Collier, Macmillan Publishers Ltd.

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

UNIT 3 RELATIONSHIP BETWEEN POPULATION AND SAMPLE

Unit Structure

3.0 Introduction

In the earlier part of this course, you came across some numerical measures that are usually employed in descriptive statistics. You will now recall that the concepts of probability learned in the last unit are said to be fundamental to the second aspect (i.e. methods of inferential statistics) which we are about to begin in this unit.

This branch of statistics is concerned with making statements about a population based on information in a sample.

In essence, it is necessary that you understand first the relationship between population and samples so as to be able to apply the methods of inferential statistics. Before discussing these points let us look at the objectives stated hereunder:

3.1 Intended Learning Outcomes (ILOs)

- Define population and sample and state the relationship between them.
- Describe the relationship between parameters and statistics and give examples of each.
- State the difference between theoretical and empirical distribution.

3.2 Main Content

3.2.1 Discussion on Population and Sample

One cornerstone of statistics is the interplay between *population and sample.* You should be aware that those numerical quantities (especially, the mean and the standard deviation) which calculated from a sample in earlier units not only describe the sample but furnish a basis for making inferences about the characteristics of the whole population.

You therefore need to be aware that a *population is a set of all measurements of interest whereas a sample is a part of the population*. For instance, a population may be the ages of all nurses in a Teaching Hospital or the number of university students.

It therefore follows that in the conduct of a statistical investigation, your initial step is to specify the common characteristics that define the population under study. You then distinguish between the target populations (i.e. population to be investigated) and population sampled (that populations about which conclusions are drawn).

You will understand these terms more suitably with the following illustration:

Example 3.1

An investigation is to determine the nutritional habits of persons over 50 years in the low-density areas of Nigerian cities. 400 persons over 50 years old who live in low-density areas of Lagos were interviewed.

In this example, the target population is all residents of low-density areas in Nigerian cities who are over 50 years old. The population sampled is all resident of low-density areas of Lagos who are over 50 years.

After the distinction between the two types of population, your next step is to generalize your results on the population sampled to the target population. You need to be cautious at this stage because the generalization may be open to controversy.

In view of this, you must be assured that the characteristics of both the population sampled and the target populations are identical. Recall that in a previous study unit it was mentioned that a statistical study of an entire population is usually impossible because of the following reasons:

(i) Size of the entire population makes the study impractical;
(ii) Cost of such study is prohibitive;
(iii) All members of the population may not be observable.

You will also recall from units 2 and 3 that the aim of selecting a sample is to ensure that the observations are unbiased. This is done through random sampling. Remember that we have defined and examined these procedures in the aforementioned study units.

We next define useful terms employed in inferential statistics.

3.2.2 Parameters and Statistics

A process by which you draw conclusions about a population given the information contained in a sample is referred to as *statistical inference.*
You will recall that most common population values for which information is required are the mean and the standard deviation. It therefore follows that these measures are usually computed from the entire population and are called parameters while computed from a set of sample measurements, they are referred to as *statistics*.

You need to associate the parameter *p* to the mean of a population and the statistics *X* to the mean of a sample. Similarly, you associate the parameter *a* to the population standard deviation and the statistics to the sample standard deviation.

Another important point you have to note in the conduct of a research is, that the value of the population parameters is virtually unknown since a part of the population is observed.

Hence the sample statistics is employed in estimating the population parameter.

Let us illustrate the relationship between population parameter and sample statistic with the following example:

A statistical investigation is to determine the mean age of nursing students in psychiatric hospitals. A random sample of 100 nursing students in these hospitals was selected and their mean age computed.

Here, the parameter of interest is the mean age (μ) of all nursing students in psychiatric hospitals. The statistics computed from a sample of 100 nursing students is the sample mean ($\overline{x}$), x estimates the true mean μ.

Recall that sample statistics estimate population parameters and as such the two sets of values are not exactly equal. This introduces sampling errors which are significant in statistical decision-making process.

3.2.3 Types of Distribution

Distributions may be classified into two general patterns. An empirical distribution is one obtained from tabulating a frequency distribution from a set of sample measurements.

On the other hand, the frequency for all the measurements in an entire population is said to be a theoretical distribution. Again, a theoretical distribution is rarely obtainable in practice because it is impractical to measure all elements in the population. Hence an empirical frequency distribution for the sample measurements approximates the theoretical frequency distribution of all population measurements.

You should be aware that knowledge of the properties of the theoretical distribution of population measurements is of crucial importance only in making inferences about a population based on sample measurements. One significant characteristic of a theoretical frequency distribution is the nature of the variable of interest in a study (i.e. is it discrete or continuous?).

In the next unit, you will learn the properties of one of the most frequently used continuous theoretical frequency distributions.

3.3 Self-Assessment Exercise(s)

An investigator studied anxiety in mothers who visited their sick children for the first time. Each of 20 randomly selected mothers was given a standardized anxiety exam immediately before she visited her child and the mean score was determined. Indicate the parameter/statistics of interest. In a study to determine the effect of the POP on the healing process of fracture of the femur, 40 patients in an orthopedic unit of a hospital were interviewed to assess the healing periods. The average healing periods were computed for the 40 patients. What would you regard as the parameter/statistics of interest?

3.4 Conclusion

In this unit, you have learned about the relationship between populations and samples. You also learned that in drawing conclusions about a

population, it is usual to make inferences on information based on a sample of the population.

Towards this direction measures computed for a sample are used to estimate by those for the whole population. This process introduces sampling error, which plays an important role in decision-making.

You also learned that there is a distinction between empirical and theoretical distributions. The former is only essential for descriptive purposes whereas the latter has several implications for statistical inferences. One of the most significant theoretical distributions is that which is continuous.

3.5 Summary

The critical concepts you learned in this units are:

- Numerical measures calculated for a sample
  a. may describe the sample itself
  b. may furnish inferences about the characteristics of the population based on information collected from the sample.
- Target population is that about which inferences are made.
- Population sampled is that from which the sample was taken.
- Generalization about the population sampled to the target population is subjective except characteristic of the two being identical is assured.
- Empirical distribution is necessary for descriptive statistics whereas theoretical distribution is essential in inferential statistics.
- Parameter is a numerical quantity calculated from the entire population e.g. population mean is denoted by p while population standard deviation is denoted by a: they are rarely obtainable in practice
- Statistics is a numerical quantity calculated from a set of sample measurements.
  $\overline{x}$ denotes sample mean; and
  S denotes sample standard deviation.

These quantities estimate the corresponding population parameter e.g. $\overline{x}$ estimates μ and s estimates σ.

3.6 References/Further Readings

Jefferies, P.M.(1995) *Mathematics in Nursing,* 4th edition, Bailliere Tindall, London: Cassell and Collier, Macmillan Publishers Ltd.

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

UNIT 4 NORMAL DISTRIBUTION

Unit Structure

4.0 Introduction

In this unit, you will learn the use of properties and applications of the most important continuous probability distribution. This is the so-called bell-shaped curve or the *normal distribution*.

It arises quite frequently in theory and practice. For instance, many physical measurements have distributions that are bell-shaped and thus it is often suitable to employ the normal distribution as that of a population or random variable.

The normal distribution is equally useful in making inferences about the mean and standard deviation of a population.

4.1 Intended Learning Outcomes (ILOs)

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

- List the properties of the normal distribution
- State the standard probabilities associated with the normal curve.
- Determine the probabilities associated with the normal curve using the standard normal table.

4.2 Main Content

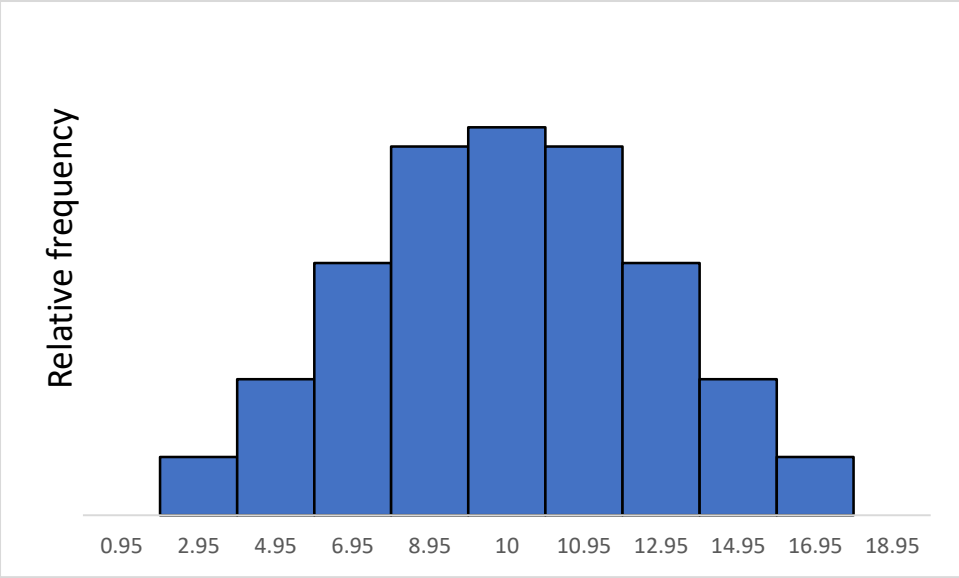
4.2.1 Discussion on the Normal Curve

You should observe that a variety of populations and random variables around us are intrinsically different. However, some like height, weight blood pressure etc. share an important characteristic. This is that the probabilities associated with them are at least approximately equal to areas under a normal curve.

From previous units, you can recall that the center points or location of a normal curve depends on the value of the mean and the variability of observations depends on the value of the standard deviation. You will now acquaint yourself with other properties of the normal curve as stated in the next part of the unit as well in the accompanying diagrams.

4.2.2 Properties of the Normal Curve

Let us consider the characteristics of the normal curve drawn below:



Birth weights (In Pounds)

Figure 4.1: ***Frequency Distribution of Birth Weights of Infants with Genetic Disorder***

You can observe from Figure 14.1 that the curve is bell-shaped about its center-point i.e., the mean. You can also observe that values that are very large or very small in relation to the mean occur infrequently. Rather the closer the values are to the mean, the frequently they occur.

You will also notice that the center point of a normal curve depends on the values of the mean and that the variability of the values depends on the values of the standard deviation. For distribution of measurements with a small standard deviation, the curve is tall and thin while for distributions with large standard deviation, the curve is short and fat (see figure 14.2)

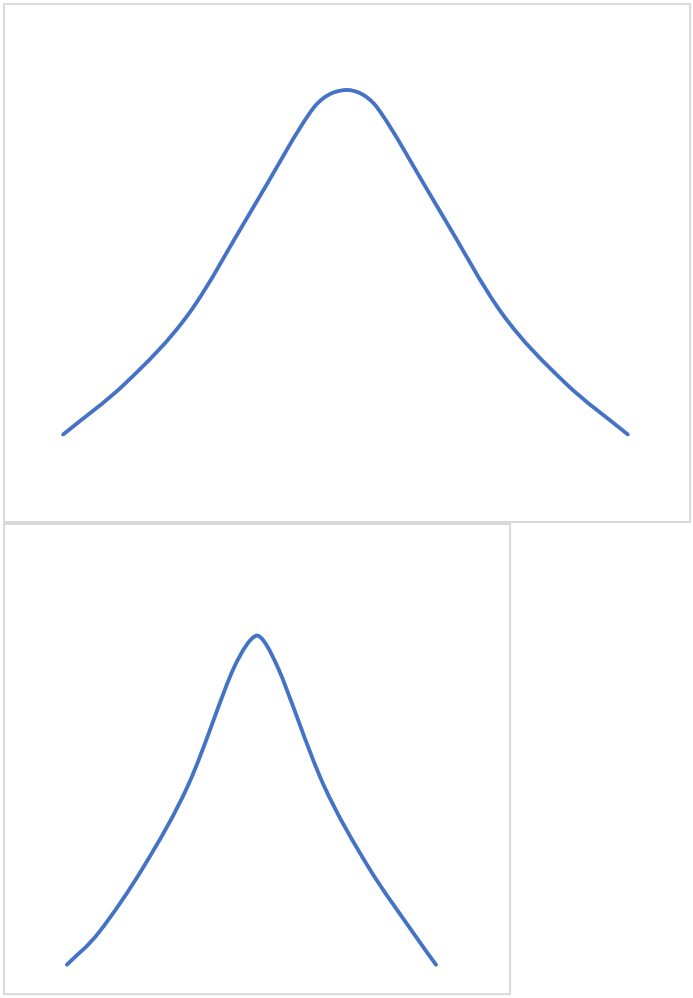


Figure 4.2: ***illustration of Normal Curves with small and large standard deviation.***

You will also observe that the normal curves of two populations with the same means but different standard deviations have different shapes as shown in figure 14.3

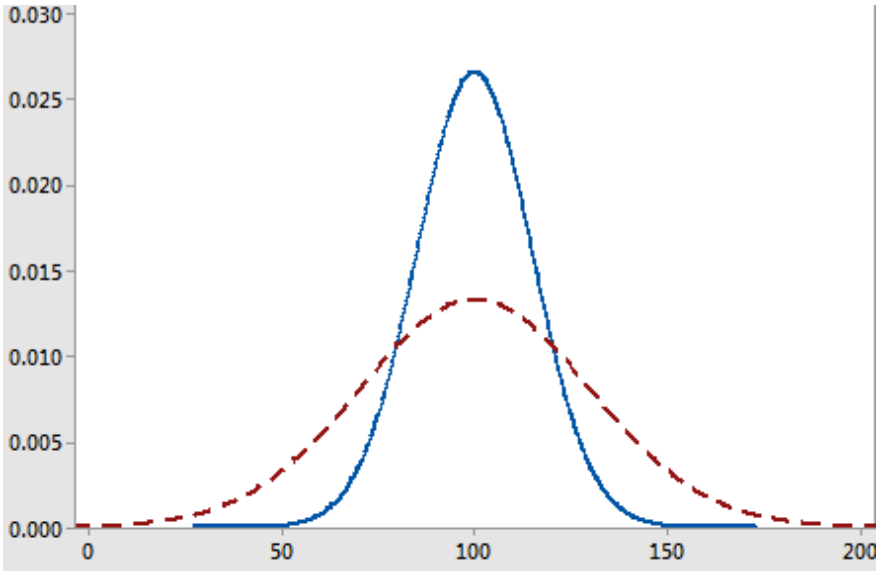


Figure 4.3: illustration of Normal Curves with the same Mean but Different Standard Deviation

You need to be aware that for a population of measurements which is normally distributed, the frequency distribution or normal curve can be easily drawn. In addition, you should be aware that for this population frequency distribution, the area under the smooth curve between two points is the probability that an observation from the population will fall in the interval specified by the two points.

Such areas of the smooth curve are obtainable from a table of normal distribution values given at the end of this study unit. The procedure will be illustrated in the next part of this unit.

4.2.3 Standard Normal Curve

There are infinitely many normal curves. One particular normal curve is a standard normal curve of z-curve. The areas under normal curves can be found once we know how to determine areas under the standard normal curve.

The horizontal axis under the standard normal curve is labeled with the letter z. the standard normal curve is shown in figure 14.4 and some of its more important properties are:

(i) The total area under the standard normal curve is equal to 1
(ii) The standard normal curve extends indefinitely in both direction approaching but never touching the horizontal axis.
(iii) The standard normal curve is symmetric about 0.
(iv) Most of the areas under the standard normal curve lies between -3 and 3

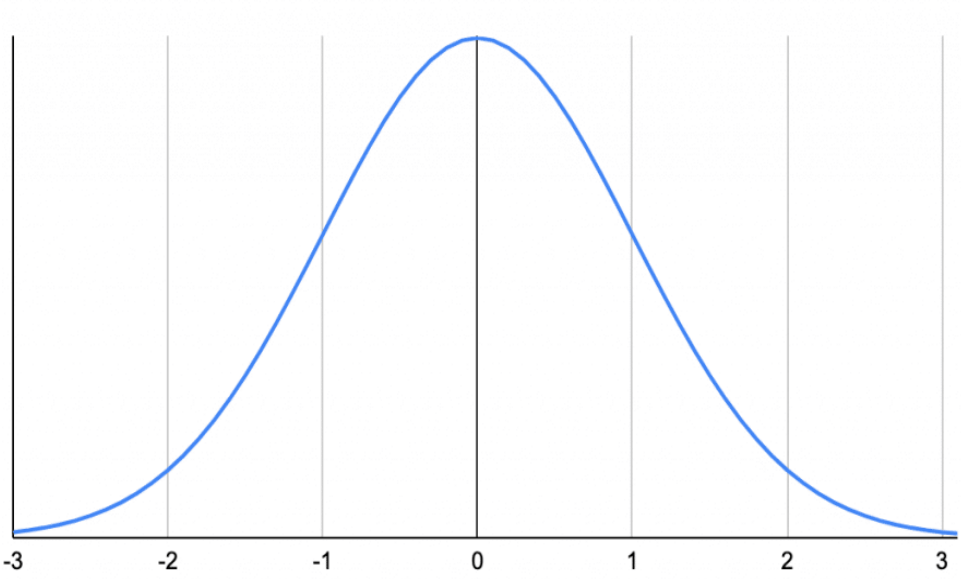


Figure 4.4: The Standard Normal Curve

4.2.4 Using the Standard Normal Table

Tables of areas under the standard normal curves have been constructed because of the importance. This table is at the end of this unit and it consists of four decimal place numbers in the body of the table. These numbers give the areas under the standard normal curve and lies to the left of a given value of z. the left page of the table is for negative values of z while the right page is for positive values of z.
The following example explains its use:

Example 4.1

Find the area under the standard normal curve that lies to the left of 1.32

Answer

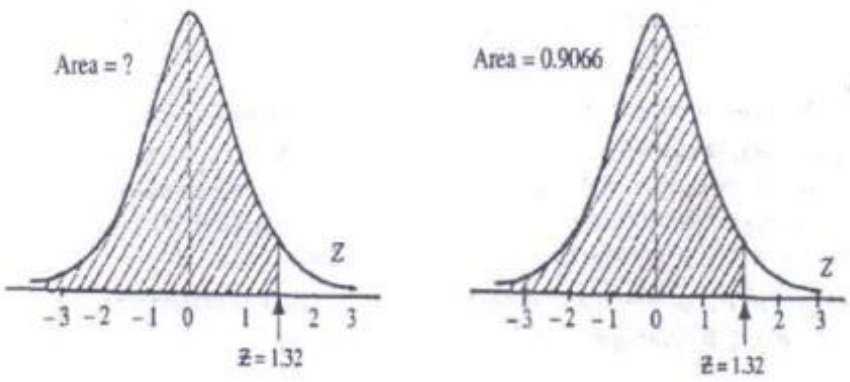


Figure 4.5 *determining the area under the standard normal curve to the left of 1.32*

From the table on the right (1.32 is positive) you go to left hand column labeled z to 1.3. then from there go across that row until you are under 0.02 in the top row. The number in the body of the table there is 0.9066. This is the area (shaded in the diagram) under standard normal curve that lies to the left of 1.32.

You have just seen one use of the Standard Normal Table. Two other important uses of that table are finding the area to the right of a given value of z and finding the area between two given values of z. we illustrate these other uses of the tables in the following examples:

Example 4.2

Find the area under the standard normal curve that lies to the right of 0.85

Answer

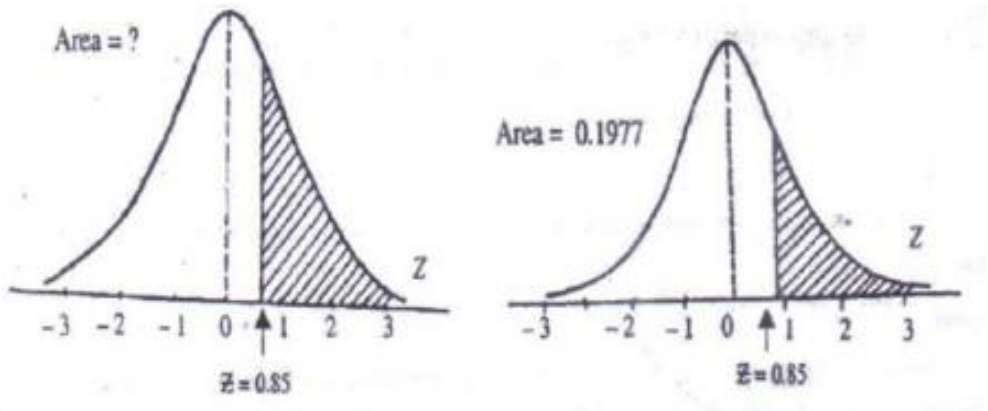


Figure 4.6 *determining the area under the standard normal curve to the right of0.85*

Since the total area under the standard normal curve is 1, the area to the right of 0.85 equals 1 minus the area to the left of 0.85. You will go down the left-hand column labeled z to 0.8. Next, go across that row until you are under 0.05 in the top row. The number in the body of the table there is 0.8023. This is the area under the standard normal curve that lies to the left of 0.85. Hence the area under the standard normal curve that lies to the right of 0.85 is 1-0.8023=0.1977.

Example 4.3
Find the area under the standard normal curve that lies between -0.57 and 1.63.

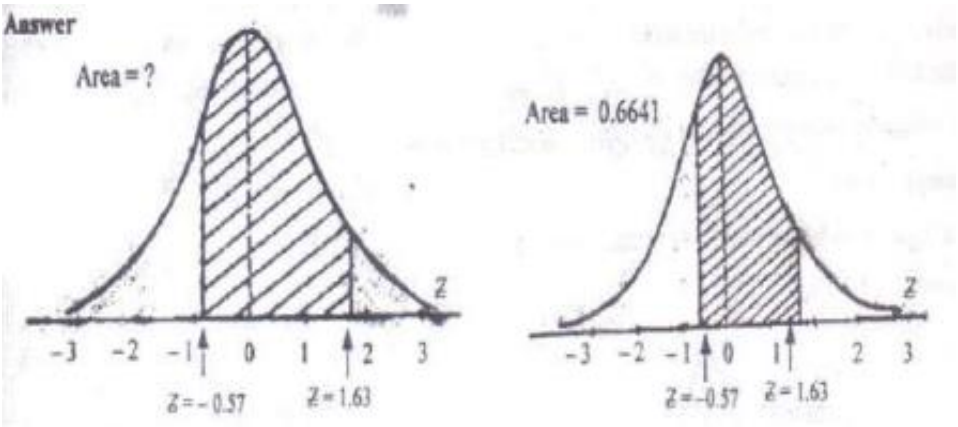


Figure 4.7: *Determining the area under the standard normal curve that lies between -0.57 and 1.63*

The area under the standard normal curve that lies between -0.57 and 1.63 equals the area to the left of 1.63 minus the area to the left of -0.57. the area to the left of 1.63 is 0.9484 and that to the left of -0.57 is 0.2843 (you will obtain this area from the left page of the table since -0.57 is negative). The area under the standard normal curve that lies between -0.57 and 1.63 is 0.9484-0.2843=0.6641.

You can also find the z-value(s) corresponding to a specified area under that standard normal curve by simply reversing the steps taken in examples 14.

4.2.5 Finding Areas under Normal Curve Using the Standard Normal Table.

It is important to be able to obtain areas under normal curves since these areas correspond to probabilities for many populations and random variables. You will recall that each normal curve is defined by two parameters namely μ and σ. μ is an indicator of the center point of the normal curve while σ indicates its shape or spread.

You will also recall that two normal curves that have the same μ parameter are centered at the same place while two normal curves that have the same μ parameter will have the same shape.

We now look at some of the most important properties of the normal curve. They are as follows:
(i) The total area under a normal curve is 1.
(ii) The normal curve extends indefinitely in both directions approaching but never touching the horizontal axis.

(iii) The normal curve with parameters μ and α is symmetric about μ.
(iv) Most of the area under the normal curve with parameters μ and σ lies between μ-3σ and μ+3σ.

We now proceed with how to find areas under the normal curves.

Example 4.4

Determine the area under the normal curve with parameters μ=5 and σ = 2 that lies
(i) To right of 6;
(ii) Between 2 and 7.

Answer

(i) The first step is for you to sketch the normal curve with parameters, μ=5 and σ =2 and shade the area to the right of 6. You will label the horizontal axis for the normal curve as x and that for the standard normal curve as z.

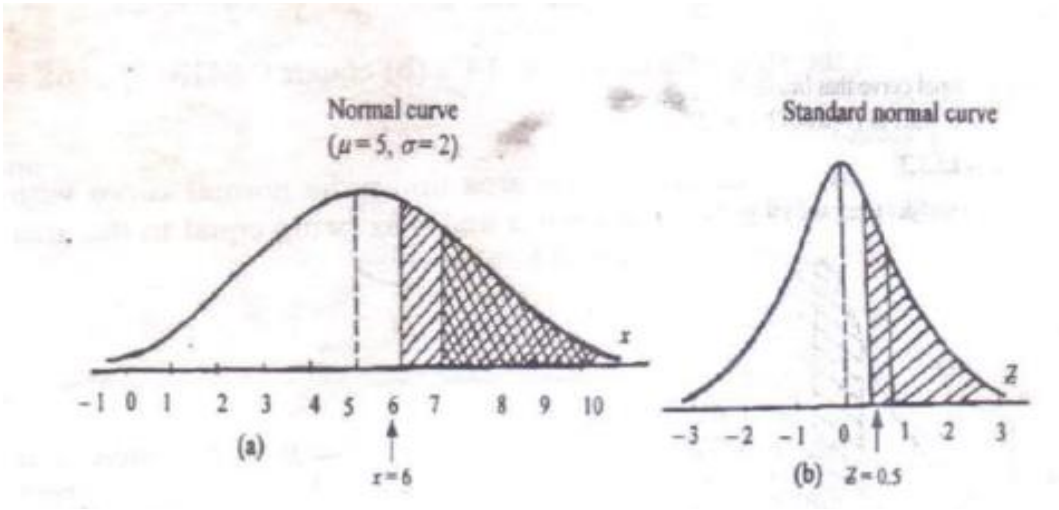


Figure 4.8

(a) Area under the normal curve with parameters, μ=5 and σ =2 that lies to the right of 6.
(b) Area under the standard normal curve that lies to the right of 0.5

You will now obtain the area under the normal curve by converting x-values to z-values by first subtracting μ and dividing σ. This conversion process is referred to as standardizing. You then obtain in (i)

$$Z = \frac{x - \mu}{\sigma} = \frac{6 - 5}{2} = \frac{1}{2}$$

i.e. areas in the shaded portion in 14.8(a) and 14.8(b) are equal. The area shaded in figure 14.8(b) is 1 - 0.6915 = 0.3085. Hence the area shaded in 14.8(a) is also 0.3085. This is the area under the normal curve with parameters μ=5 and σ =2 that lies to the right of 6 is 0.3085.

(ii) In this case you need to find the area under the normal curve with parameter μ=5 and σ =2 that lies between 2 and 7 as shown in figure 14.9(a)

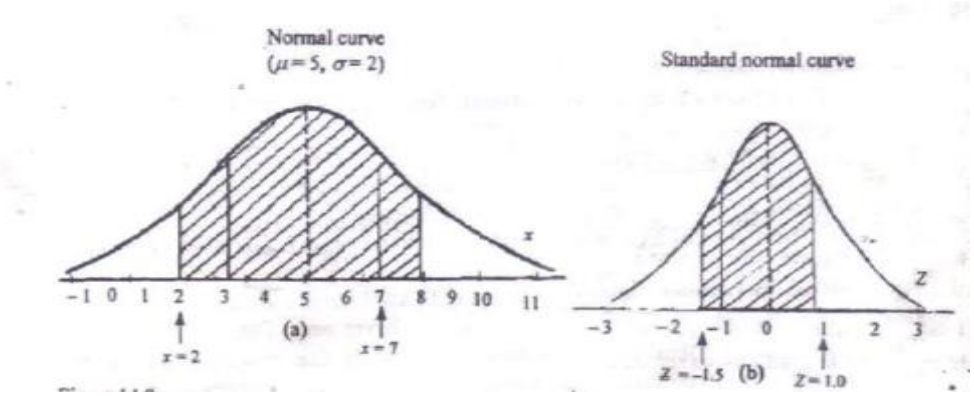


Figure 4.9

(a) Area under the normal curve with parameters, μ=5 and σ =2 that lies between 2 and 7.

(b) Area under the normal curve with parameters, μ=5 and σ =2 that lies between x = 2 and x = 7 is equal to the area under the standard normal curve that lies between

$Z = \frac{x-\mu}{\sigma} = \frac{2-5}{2} = -1.5$ and $Z = \frac{x-\mu}{\sigma} = \frac{7-5}{2} = 1.0$

Using the table, the shaded area in Fig.14.9 (b) equals 0.8413 – 0.0668 = 0.7745

In summary, you will determine the area under the normal curve with parameters μ and σ that lies between a and b as being equal to the area under the standard normal curve that lies between

$\frac{(a-\mu)}{\sigma}$ and $\frac{(b-\mu)}{\sigma}$

Also you will find the area under the normal curve with parameters μ and σ that lies to the right (or left) of a particular x-value by first converting to the z-score and then using the Standard Normal table to find the area under the standard normal curve that lies to the right (or left) of the z-score.

4.3 Self-Assessment Exercise(s)

1. In a hypertensive screening program, health workers determined that the mean systolic blood pressures in a social class is 142 mm/Hg with a standard deviation σ of 10 mm/Hg. A decision was taken that any individual in this class who had a systolic

blood pressure for standard deviations above the mean normal systolic pressure would be considered hypertensive. A female from this group was tested and found to have a systolic blood pressure of 172mm/Hg. How would she be classified?

4.4 Conclusion

In this unit, you learned that frequency distributions may be continuous or discontinuous, symmetrical or skew. Also you learned that a particular shape of continuous symmetric distribution is known as the normal distribution which has very important properties for statistical analysis. Other shapes of distribution can often be transformed mathematically to appropriate normality.

4.5 Summary

In this unit, the critical concepts learned include the following:

- Normal distribution

(a) Is a continuous theoretical frequency distribution and has the following characteristics:
(i) It is bell shaped
(ii) It has continuous measurements; and
(iii) It is symmetric around its mean (μ)
(b) It has its location of centre determined by μ and spread by σ
(i) For small values of σ curve is tall and thin;
(ii) For large values of σ curve is short and fat.

- To determined probabilities associated with the normal curve, the raw scores x are first converted to standard scores z by the formula $Z = \frac{x-\mu}{\sigma}$
- Areas under standard normal curve are found in tables.

4.6 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

Weiss, A.L (1996) *Elementary Statistics,* 3rd edition, Addison-Wesley Publishing Co. Inc.

UNIT 5 SAMPLING DISTRIBUTION OF THE MEAN AND THE CENTRAL LIMIT THEOREM

Unit Structure

5.0 Introduction

In the preceding units, you have studied sampling, descriptive statistics, probability, random variables and the normal distribution. Now you will need to learn that these diverse topics can be integrated to serve as a foundation for inferential statistics.

You will first understand why sampling distribution of the mean is of unique importance in statistics. Next, you will learn the properties of the mean.

One of these properties lead on to the Central Limit Theorem. Before discussing these concepts let us look at the objectives of this study unit stated hereunder:

5.1 Intended Learning Outcomes (ILOs)

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

- Describe in words the meaning of sampling error
- Explain concisely the meaning of a derived distribution
- Explain the steps required to generate a sampling distribution
- State the Central Limit Theorem.

5.2 Main Content

5.2.1 Sampling Error: The Need for Sampling Distribution

Recall that using a sample to acquire information about a population is often preferable to taking a census, in which data for the entire population are collected.

However, you are aware that since a sample from a population provides data for only a portion of the entire population, it is unlikely the sample yields perfectly accurate information about the population.

Hence you need to anticipate an amount of error in the sampling procedure. This error is called sampling error, in order to answer questions about the accuracy of estimating a population mean by a sample mean, there is the need to know the distribution of all possible sample means that could be obtained.

This is a derived distribution from the original parent distribution and it is called the *sampling distribution of the mean.* In the succeeding parts of this unit, we will discuss how to generate a sampling distribution and study its properties.

5.2.2 Derived Distribution-Generation

The sampling distribution of the mean is of unique importance in statistics. So as to gain insight into its meaning and usefulness, let us consider how to generate the distribution of sample means for a given sample of size *n*

You need to note that a sampling distribution is rarely generated in practice rather you only need to have knowledge of its properties so as to be able to make statistical inferences.

To generate a sampling distribution of Y's you proceed as follows:

(i) Construct a population by recording population values on slips of paper and placing the slips in a container:
(ii) Compute the mean Y of a sample of size *n* selected from the population.
(iii) Replace these *n* observations in the population.
(iv) Repeat steps 2 and 3 until a large number of samples of size *n* have been drawn.
(v) The resulting frequency distribution of each of the *X's* approximates the distribution of the sample mean values or the sampling distribution of means for a given *n.*

You should note that the process of generating a complete sampling distribution is an impossible task for parent populations with few observations. We now state the properties of the sampling distribution of the means.

5.2.3 Properties of the Distribution of Sample Means (or Sampling Distribution of the Mean)

These are stated as follows:

1. The mean, $\bar{X}$ of the distribution is equal to μ, the mean of the parent population from which the samples are drawn.
2. The standard deviation δ_x of the sampling distribution of $\bar{X}$ is equal to $\frac{\delta}{\int n}$ the standard deviation of the parent population divided by the square root of the number in each sample. The *standard deviation of X is called the standard error of the mean* and is given in terms of the standard deviation of the individual values of the parent distribution.

It allows the computation of an estimate of the standard deviation of mean values from a single sample of individual observations.

3. When sampling is from a parent population whose distribution is normal, the shape of the sampling distribution is that of the normal curve. The sampling distribution is approximately normal when sampling is from non-normal parent population. Approximation to normality becomes closer and closer as the sample size increases.

5.2.4 Central Limit Theorem

The third property of the sampling distribution of the mean ($\bar{X}$) is an extra-ordinary fact called the *central limit theorem.* It states that: *for any given population, not necessarily normally distributed and having mean μ and standard deviation σ the sampling distribution for fixed n which is generated from this population will be approximately normally distributed with mean μ and standard deviation* $\frac{\sigma}{\int n}$ *.*

You should note that this assumption of normality of $\bar{X}$ is the cornerstone of statistical inference. We now illustrate the use of the Central Limit Theorem with the following example:

Example 5.1

In a hypertensive survey, it was determined that the mean systolic blood pressure for normal males in social class is 120 mm/Hg with standard deviation of 10 mmHg. Describe the frequency distribution of all

possible sample means of size 625 from the given population of systolic blood pressure measurements.

Answer

The parent population of systolic blood pressure measurements has a mean equal to 120 mm/Hg and a standard deviation σ equal to 10 mm/Hg. Applying the properties of sampling distributions and the Central Limit Theorem, the population of sample mean $(\overline{X})$ of size n=625 that could be drawn from this parent population is normally distributed with mean μ_x equal to the mean of the parent population and standard deviation $\sigma_{\bar{X}}$ to the standard deviation σ of the parent population divided by the square root of the sample size.
Therefore,

$$\mu_{\bar{X}} = \mu = 120mm/Hg$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{625}} = \frac{10}{25} = 0.4mmHg$$

5.3 Self-Assessment Exercise(s)

Let the population of interest consists of the heights of five nursing students in a hospital unit. If the heights (in inches) are 76,78,79,81,86

(i) Determine the sampling distribution of the mean for random samples of two heights from the population of five heights.

(ii) Indicate your observations about sampling error when the mean, $\bar{X}$ of a random sample of two heights is used to estimate the population mean, μ

(iii) Employ the sampling distribution of the mean obtained in (i) to find the probability that the sampling error made in estimating the population mean μ, by the mean, $\bar{X}$ of a random sample of two heights will be at most 1 inch; that is, for a random sample of size two, determine the probability that $\bar{X}$ will be within 1 inch of μ.

5.4 Conclusion

In this unit, you studied one of the most important ground work of statistical inference. This is the sampling distribution of the mean. Its properties were suitably explained to you of which results in the Central Limit Theorem was applied to a problem.

5.5 Summary

The following critical concepts were studied in this unit:

- Sampling error is the failure of a statistic $\bar{X}$ to be exactly equal to its corresponding population parameter.
- If many samples are drawn from a parent population and the statistic $\bar{X}$ computed for each sample, the values of $\bar{X}$ will cluster around the true value of the parameter, μ
- The Central Limit Theorem states that for a relatively large sample size, the random variable, $\bar{X}$ is approximately normally distributed has mean μ_x and standard deviation $\sigma_\mu = \sigma/\sqrt{n}$ regardless of the population's distribution. The approximations become better and better with increasing sample size.
- The standard deviation of the population of $\bar{X}$ is $\sigma/\sqrt{n}$'s the standard error of the mean.

5.6 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

Weiss, A.L (1996) *Elementary Statistics,* 3rd edition, Addison-Wesley Publishing Co. Inc.

MODULE 4 HYPOTHESIS TEST

Introduction

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process.

UNIT 1 MEAN ESTIMATION

1.0 Introduction

In this unit, you will begin the study of inferential statistics. This consists of a set of procedures used to draw conclusions about a large

body of data called a population based on a subset of data from the population i.e., a sample.

There are two main areas of statistical inference namely estimation and testing of hypothesis. In this unit you will examine methods for estimating the mean of a population.

1.1 Intended Learning Outcomes (ILOs)

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

- Calculate and interpret confidence intervals on a single population mean *m* for

(i) A population with known standard deviation σ,
(ii) A population with unknown standard deviation σ, and

- Calculate and interpret confidence intervals on the difference in two population means, $\mu_1 - \mu_2$ for

(i) Population standard deviation known;
(ii) Population standard deviation unknown; and

- Calculate and interpret confidence intervals on the true mean difference D in a paired experiment.

1.2 Main Content

1.2.2 Estimating a Population Mean

A common problem in statistics is that of estimating a population mean *p*. for instance, we might want to know

(i) The mean petrol kilometer of a new-model car;
(ii) The mean weight gain for patients suffering from a certain debilitating disease who are on a diet.
(iii) The mean tar content of a certain brand of cigarette.

For a small population, you can ordinarily determine μ exactly by census-taking and then calculating μ from the population data. But for a large population, census taking is impractical, very expensive or impossible.

Nonetheless, you can obtain fairly accurate information about μ by taking a sample from the population.
You may recall the estimation of the mean of the number of geriatric patients in five urban hospitals in Nigeria in Example 8.2.2.1. The mean obtained is

$$\overline{x} = \sum \frac{x_i}{n} = \frac{145}{5} = 29$$

An estimate of this kind, which is based on the sample data is called a point estimate for μ because it consists of a single point.

You need to recall from the last study unit that it is unreasonable to expect that a sample mean $\overline{x}$ will exactly equal the population mean μ. You should anticipate some sampling error. It is therefore necessary that in addition to reporting a point estimate of μ, you need to furnish information on the accuracy of the estimate.

You can do this by providing a confidence interval estimate for μ. When you furnish a confidence interval estimate of μ, you then use the mean of the sample to construct an interval of numbers and express how confident you are that μ lies in that interval.

Before seeing how to obtain such confidence intervals, let us formally define the new terminologies, which incidentally apply to other numerical summaries apart from the mean (e.g the standard deviation).
A confidence interval estimate for a parameter comprises of an interval of numbers obtained from a point estimate of the parameter together with a percentage that specifies how confident we are that the parameter lies in the intervals. The confidence percentage is referred to as the *confidence level*.

You will now learn a technique for obtaining a confidence interval for a population means. Before proceeding, you will need to recall the following Key Facts:

(i) The Central Limit Theorem;
(ii) The Sampling Distribution of the Mean for General Population; and
(iii) The empirical rule for normally distributed random variables which states as follows: for any normally distributed random variable x:

Property 1: The probability is 0.6826 that an observed value of x will lie within one standard deviation to either side of the mean;

$$P(\mu_x - \sigma_x < x < \mu_x + \sigma_x) = 0.6826$$

Property 2: The probability is 0.9544 that an observed value of x will lie within two standard deviations to either side of the mean;

$$P(\mu_x - 2\sigma_x < x < \mu_x + 2\sigma_x) = 0.9544$$

Property 3: The probability is 0.9974 that an observed value of x will lie within three standard deviations to either side of the mean

$$P(\mu_x - 3\sigma_x < x < \mu_x + 3\sigma_x) = 0.9974$$

Let us look at an example

Example 1.1

Let a random sample of size n be taken from a population with mean μ and standard deviation σ. In addition, let either the population be normally distributed or the sample size be large ($n \geq 30$).
Determine

(i) the probability that the interval from $x - 2.\frac{\sigma}{\sqrt{n}}$ to $x + 2.\frac{\sigma}{\sqrt{n}}$ will contain the population mean μ.

(ii) Interpret the result in (i) using percentages.

Answer

$\overline{x}$ is normally distributed since it is assumed that the population is normally distributed or sample size is large. Hence by property 2 earlier mentioned, the probability is 0.9544 that an observed value of $\overline{x}$ lies within two standard deviations to either side of the mean:

$i.e.$ $P(\mu_x - 2\sigma_x < x < \mu_x + 2\sigma_x) = 0.9544$

But

$$\mu_{\bar{X}} = \mu \text{ and } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

The last equation can be rewritten as

$$P(\mu - \frac{2\sigma}{\sqrt{n}} < \overline{x} < u + \frac{2\sigma}{\sqrt{n}}) = 0.9544$$

Stated in words, the last equation means that the probability is 0.9544 that $\overline{x}$ will lie within $2.\frac{\sigma}{\sqrt{n}}$ of μ. It also means that u will lie within $2.\frac{\sigma}{\sqrt{n}}$ of $\overline{x}$.

Hence the last can be rewritten as:

$$P\left(\overline{x} - 2.\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 2.\frac{\sigma}{\sqrt{n}}\right) = 0.9544$$

From whence we have that the population that the interval from $\overline{x} - 2.\frac{\sigma}{\sqrt{n}}$ to $\overline{x} + 2.\frac{\sigma}{\sqrt{n}}$
Will contain μ is 0.9544

You need to realize that the random variable is $\overline{x}$ not μ. In addition, you should note that the population mean μ is a fixed number, although it may be unknown, the sample size $\overline{x}$ is a random variable and its value depends on chance i.e. on which the sample is obtained

(ii). One interpretation is that about 95.44% of all samples of size n have the property that the interval with endpoints $x + 2.\frac{\sigma}{\sqrt{n}}$ contains the population mean μ. Another interpretation is that if you take a large

number of random samples of size n, then about 95.44% of the samples obtained will possess the property that the interval with endpoints $x \pm 2.\frac{\sigma}{\sqrt{n}}$ contains the population mean μ.

1.2.3 Interpretation of Confidence Intervals on μ.

Confidence intervals may be interpreted in two ways. One of the ways is theoretical in nature and its consideration is long term. For instance, if a series of samples were obtained with a construction of 95% confidence intervals for each sample on the true mean μ, then ultimately, the relative frequency with which the calculated intervals actually cover μ is 95%.

The second interpretation specifies 95% confidence in the interval $\overline{x} + 1.96\,\sigma/\sqrt{n}$ for a particular sample of the true population mean μ.

1.2.4 Confidence Interval for one Population Mean when σ in Unknown

In the preceding sections of this study unit, you learned how to determine the confidence interval for a population mean μ when the standard deviation σ is known.

The procedure is based on this fact:

If whenever the population being sampled is normally distributed or the sample size is large, then the random variable $\overline{x}$ is normally distributed (or approximately so) and has mean $\mu_x=\mu$ with standard deviation,

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

The standardized random variable $z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$

Has the standard normal distribution.

But in practice, the population standard deviation is unknown and as such you cannot base your confidence-interval procedure on $\overline{x}$ (*and thus z*). However, since the sample standard deviation *s* is a point estimate of the population standard deviation, σ by *s* in

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

and base your confidence-interval method on the resulting random variable. Your next task is to identify the probability distribution of this new random variable i.e.

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

In carrying out this task, you will assume that the population being sampled is normally distributed. The obvious problem with the substitution of s for σ is that in addition to the variability of $\overline{x}$, s now varies with each sample. This additional variability in s is taken into account by use of the Student t-distribution.

This distribution is *symmetric, bell-shaped* and *centered* on a mean of O. furthermore, the exact shape of the t-distribution depends on a value called *degrees of freedom. Degrees of freedom for the t-distribution is defined to be n-1 for the case of estimating a single population mean μ.* You need to know that the concept of degrees of freedom is used for the *t*-distribution in obtaining the confidence coefficient from the table of *t*-values.

The rationale, procedure and interpretation of confidence intervals are the same for both the *t*- and *z*- distributions. For a confidence interval on μ using t, the estimate s is substituted for σ and an appropriate value for t is obtained from a table of t values. The formula is

$$\overline{x} \pm t(s/\sqrt{n}$$

Let us illustrate a use of *t*-table with the following example

Example 1.2

For *t*-curve with 13 degrees of freedom, determine $t_{0.025}$ i.e. find the *t*-value having area 0.02 to its right.

Answer

To find the *t*-value in question we use the Table whose portion has been repeated here for reference:

Table 16 Value of t_a

df	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$	Df
-						-
-						-
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	-	-	-	-	-	-

To locate $t_{0.025}$ with $df = 13$, you will go down the outside column labeled df to 13 and then go across that row until we are under the column headed $t_{0.025.}$ The number in the body of this table is 2.160 which is the required value.

1.2.5 Obtaining Confidence Intervals for a Population Mean when σ is Unknown

Having discussed t-distribution and t-curves, you can now see how to develop a method to obtain a confidence interval for a population mean, μ when a population being sampled is normally distributed and the population standard deviation is unknown. The procedure is as follows:
Let a random sample of size *n* be taken from a normal distributed population with mean μ. Then the random variable.

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

Has the *t-distribution* with (*n-1*) degrees of freedom i.e probabilities for that random variable are equal to areas under the *t-curve* with *df=n-1*.
Therefore,

$$P\left(-t_{a/2} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{a/2}\right) = 1 - a$$

This equation may be rewritten as

$$P\left(\bar{x} - t_{a/2}.\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{a/2}.\frac{s}{\sqrt{n}}\right) = 1 - a$$

From whence you have that once the sample is taken, the interval from

$$\bar{x} - t_{a/2}.\frac{s}{\sqrt{n}} \;\; to \; \bar{x} + t_{a/2}.\frac{s}{\sqrt{n}}$$

Will be a (*1-a*) – level confidence interval for μ

Example 1.3

To estimate the mean gestation period of domestic dogs, 15 randomly selected dogs are observed during pregnancy. Their gestation periods (in days) are:

62.0	61.4	59.8	62.2	60.3
60.4	59.4	60.2	60.4	60.8
61.8	59.2	61.1	60.4	60.9

Obtain a 95% confidence interval for the mean gestation period μ of the domestic dogs. [this example is taken from Elementary Statistics by N.A. Weiss]

Answer

The sample size is moderate, a normal distribution plot shows no outliers and falls roughly in a straight line.

For a confidence level of *1-a*, you will use the *t*-table to find $t_{a/2}$ with *df = n-1*, where *n* is the sample size.

The specified confidence level is 0.95, so $a = 0.05$. Since $n = 15$, we have $df = n - 1 - 15 - 1 = 14$. the t-table shows that for $df = 14$

$$t_{a/2} = t_{0.05/2} = t_{0.025} = 2.145$$

The confidence interval for μ is from

$$\bar{x} - t_{a/2}.\frac{s}{\sqrt{n}} \ \ to\ \bar{x} + t_{a/2}.\frac{s}{\sqrt{n}}$$

$t_{a/2} = 2.145$. Find $\bar{x}$ and s from the data in the usual way we have $\bar{x} = 60.69$ and $s = 0.90$. Hence, a 95% confidence interval for μ is from

$$60.67 - 2.145\frac{(0.90)}{\sqrt{15}}\ to$$

$$60.69 - 2.145\frac{(0.90)}{\sqrt{15}}$$

or 60.19 to 61.18

You can then be 95% confident that the mean gestation period μ of the domestic dog is somewhere between 60.19 to 61.18 days.

1.2.6 Confidence Intervals on the Difference in Two Population Means: Population Standard Deviations Known

In the previous section of the study unit, you saw the rationale and the method ascribed to a confidence interval on a single population mean μ. In medical and nursing practice, it is usual to estimate the true difference in two population means μ_1 - μ_2, where μ_1 and μ_2 are the respectively true means for the first and second populations.

Here, you will have that μ_1 - μ_2 is the population parameter of interest while the difference in sample means $(\bar{X}_1 - \bar{X}_2)$ serves as a point estimate of the population difference μ_1 - μ_2.

You need to be aware that with a rigorous proof, it can be assumed that if possible samples of size *n*, are drawn from the first population of size N_2 and all possible samples of size n_2 are drawn from the second population of size N_2 such that from these samples all possible differences $\bar{X}_1 - \bar{X}_2$ are computed, then the frequency distribution of these differences is that of the normal distribution with mean μ_1 - μ_2 and standard deviation

$$\sqrt{(\sigma_1^2 \ln_1) + (\sigma_2^2 \ln_2)}$$

A formal statement of this fact is that the sampling distribution of the differences X_1 –X_2 of independently drawn samples is normally distributed with mean μ_1 - μ_2 and standard deviation

$$\sqrt{(\sigma_1^2 \ln_1) + (\sigma_2^2 \ln_2)}$$

You should note that if the two sampled populations are not normally distributed or if the form of the frequency distributions of the populations is unknown, the sampling distribution of the differences, $\bar{X}_1 - \bar{X}_2$ is at least approximately normally distributed with mean μ_1 - μ_2 and standard deviation $\sqrt{(\sigma_1^2 \ln_1) + (\sigma_2^2 \ln_2)}$ for large sample sizes n_1 *and* n_2.

It therefore follows from the general form of a *confidence interval* on a population parameter i.e

(Estimate of parameter) $\pm$ (Confidence Coefficient) $\times$ (Standard error of estimate) that the 95% and 99% confidence interval on the true population difference μ_1 - μ_2 are

$$(\bar{X}_1 - \bar{X}_2) \pm 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(\bar{X}_1 - \bar{X}_2) \pm 2.56 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Where $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ is the standard error of the estimate $(\bar{X}_1 - \bar{X}_2)$

When interpreted, you may be 95% (or 99%) confident that the interval spans the true differences in population means μ_1 - μ_2.

1.3 Self-Assessment Exercise(s)

1.4 Conclusion

In this unit, you learned that having observed a mean, or any other numerical summary that may be computed from a sample of observations, it is possible to give limits within which the corresponding value in the population lies, with a known degree of confidence.

You need to know that it is conventional to employ 95% confidence limits but other degrees of confidence may be used if desired.

1.5 Summary

The critical concepts learned in this study unit are as follows:

- There are two aspects of statistical inference namely: *estimation* and *hypothesis testing.*
- A method of estimating population parameters using information in the sample is the confidence interval procedure.
- A confidence interval may be interpreted in two ways:

(i) If a series of samples were obtained and 95% confidence intervals on the true parameter constructed for each sample, then ultimately, the relative frequency with which the computed intervals actually cover the true parameter is 95%.

(ii) Using a 95% confidence interval for the single sample obtained in practice, you are 95% confident that the confidence interval computed for that particular samples covers the true population parameter.

- The general form of a confidence interval is as follows: (Estimate of parameter) $\pm$ (Confidence Coefficient) $\times$ (Standard error of estimate).
- A confidence interval on the true population means μ when the population standard deviation is known is

$$\pm Z(\bar{X})\frac{\sigma}{\sqrt{n}}$$

Where z, the *confidence coefficient* is a standard normal table value.

- A confidence interval on the true population means μ when the population standard deviation σ is not known is $\bar{x} + t\frac{s}{\sqrt{n}}$, where *t*, the confidence coefficient is a value from the *t*-distribution.
- A confidence interval on the true difference in population means μ_1 - μ_2 when the population standard deviation σ_1 and σ_2 are known is

$$(\bar{X}_1 - \bar{X}_2) \pm Z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

1.6 References/Further Readings

Jefferies, P.M.(1995) *Mathematics in Nursing,* 4th edition, Bailliere Tindall, London: Cassell and Collier, Macmillan Publishers Ltd.

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

UNIT 2 FUNDAMENTALS OF HYPOTHESIS TESTS

Unit Structure

2.0 Introduction

In the preceding study unit, we study one aspect of inferential statistics. These involved methods of obtaining confidence intervals for one population mean.

This study unit and the two succeeding units will focus on how you can employ the sample means $\overline{x}$ to make decisions about the hypothesized values of a population mean μ. For instance, you might wish to use the mean success of treatment of tuberculosis for a sample of people to penultimate year to generalize for all such people last year. Such statistical inference lies in the realm of hypothesis test.

This study unit will discuss the fundamentals of hypothesis test while you will learn standard procedures employing in hypothesis testing in the next two study unit. You will need to examine the objectives for this unit stated hereunder.

2.1 Intended Learning Outcomes (ILOs)

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

- List the elements of a statistical test and describe in words what is meant by

(i) Null hypothesis;
(ii) Test statistics;
(iii) Level of significance;
(iv) Rejection region; and
(v) Decision or conclusion.

- Recognize these elements in a given research situation.

2.2 Main Content

2.2.1 Nature of Hypothesis Testing

Statistical inference is employed methods in this decision-making process is a hypothesis test. This is simply a statement that something is true. You need to be aware that a hypothesis test involves two mutually exclusive hypotheses referred to as the *null* and the *alternative hypotheses*.

You should be aware that the problem in a hypothesis test is to decide whether or not the null hypothesis should be rejected in favor of the alternative hypothesis.

2.2.2 Choosing the Hypotheses

Your first step in a hypothesis test is to decide on the null and the alternative hypothesis. Suppose your hypothesis tests is specifically for one population mean μ.

Then you should state the null hypothesis in the form

$$H_0: \mu = \mu_0$$

The choice of your alternative hypothesis depends on and must reflect the purpose of the hypothesis test. Your choices of alternative hypothesis could be any of the three forms:

(i) $H_0: \mu \neq \mu_0$ referred to as a *two-tailed test*
(ii) $H_a: < \mu_0$ referred to as a *left-tailed test*
(iii) $H_a: > \mu_0$ referred to as a *right-tailed test*

A right or left-tailed hypothesis test is said to be one-tailed. Let us illustrate these terms with the following examples.

Example 2.1
Suppose a clinical investigator wishes to determine whether infants with Disorder *X* have the same birth weight as normal infants. The mean birth weight of all normal infants is 7 pounds. For a random sample of 25 infants, suppose the mean birth weight is calculated to be 6.2 pounds.
(i) Determine the null hypothesis for the hypothesis test.

(ii) Determine the alternative hypothesis for the hypothesis test.
(iii) Classify the hypothesis test as two-tailed, left-tailed or right-tailed.

Answer

(i) Let μ denote the mean birth weight of normal infants. The null hypothesis is H_o: μ ≠ 7
(ii) The alternative hypothesis is H_a: μ ≠ 7
(iii) The hypothesis test is two-tailed

2.2.3 Logic of Hypothesis Testing

You have learned how to choose suitable null and alternative hypotheses for a hypothesis test. The next step is how to decide on acceptance or rejection of the null hypothesis in favor of the alternative hypothesis.

To do this you need to take a random sample from the population and determine the consistency or otherwise of the sample data with the null hypothesis. If the sample data are consistent with the null hypothesis, the null hypothesis is accepted. On the other hand, if the sample data are inconsistent with the null hypothesis, the null hypothesis is rejected in favor of the alternative hypothesis.

The next example gives a precise criterion for deciding whether or not to reject the **null** hypothesis.

Example 2.2

To check whether the instrument measuring the systolic blood pressures measurement of 25 patients, the recording for these patients are as follows (normal systolic blood pressure measurement is 140mm/Hg and the population standard deviation is known to be 10 mmHg).

140	180	130	160	120
160	150	120	140	140
180	170	150	130	160
140	180	180	180	170
160	150	140	150	180

(i) State the null and alternative hypotheses.
(ii) Obtain a precise criterion for deciding whether or not to reject the null hypothesis in favor of the alternative hypothesis.
(iii) State the conclusion.

Answer

(i) H_0: μ = 140
H_0: μ ≠ 140
Test StatisticType equation here.

$$\overline{x} = \frac{\sum x_i}{n} = \frac{3860}{25} = 154.4$$

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{154.4 - 140}{10/\sqrt{25}}$$

Where $\overline{x}$ =3860, μ = 140, $n = 25$, σ = 25

(ii) The sample mean is 7.2, standard deviations above the null hypothesis mean of 140. (The assumption that the random variable $\overline{x}$ is normally distributed has been made).

(iii) We conclude that the null hypothesis be rejected in favor of the alternative hypothesis $\mu \neq 140$

2.2.3.1 Terms, Errors and Hypothesis

In order to fully understand the nature of hypothesis testing, you will need to learn more additional notions. These notions are defined and discussed so as to be able to interpret the possible conclusions of a hypothesis test in terms of them.

Recall the basis employed in Example 17.2.2.1 for deciding whether to reject the null hypothesis. You will remember that we used the random variable

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

This is called the *test statistic* for the hypothesis test. The following graph portrays the criterion used to decide on the rejection or otherwise of the null hypothesis:

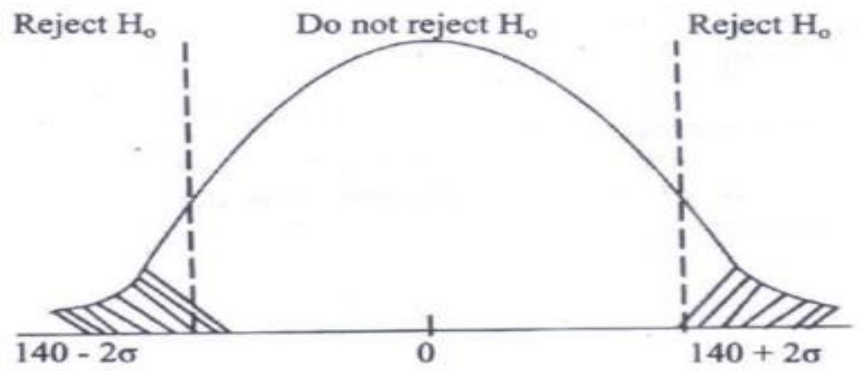


Fig 2.1

The set of values for the test statistic that leads to rejection of the null hypothesis is called the *rejection region.*

In this case, the rejection region consists of all z-values that lie either to the left of $140 - 2\sigma$ or the right of $140 + 2\sigma$ that part of the horizontal axis under the shaded area in Figure 17.1

The set of values for the test statistic that does not lead to the rejection of the null hypothesis is called the *non-rejection or acceptance region*.

The values of the test statistic that separate the rejection and non-rejection region are called the *critical values.* In the example, the critical values are $z = \pm(140 - 2\sigma)$, as shown in Figure 17.2 below:

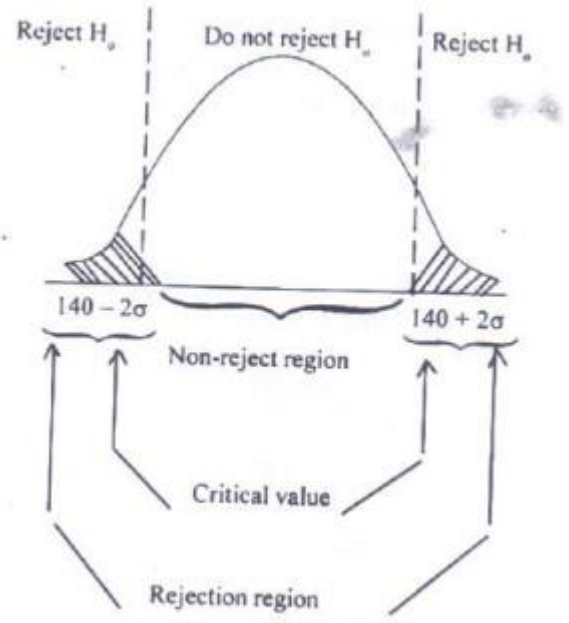


Fig 2.2

You need to be aware that for

(i) A two-tailed test, the null hypothesis will be rejected if the test statistic is either too small or too large. It therefore follows that the rejection region for such a test consists of two parts, one on the left and one on the right.

(ii) A left-tailed test, the null hypothesis will be rejected only if the test statistic is too small. Thus, the rejection region for such a test consists of the part on the left.

(iii) A right-tailed test, the null hypothesis will be rejected only if the test statistic is too large. Thus, the rejection region for such a test consists of that part on the right.

2.2.3.2 Type I and Type II Errors

One important point you need to note is that whenever statistical inference procedures are employed, it is possible to reach incorrect result. This is because a subset of the parent population is used to draw conclusion.

In view of this, four possible outcomes arise from hypothesis testing. You need to be aware that two of the possible outcomes lead to incorrect decisions. The outcomes are portrayed in Table 17.2 below:

	True	False
Accept H_0	Correct Decision	Type II error
Reject H_0	Type I error	Correct Decision

Table 2.2 : Possible outcomes of a Hypothesis Test

You need to notice that type I error result from rejecting the null hypothesis when it is infected true; while type II error results from not rejecting the null hypothesis when it is infected false.

Another point you must also observe is that the null and the alternative hypotheses in a hypothesis test are exhaustive. That is, if the null hypothesis is false the alternative hypothesis is true and vice versa.

2.2.3.3 Probabilities of Type I and Type II Errors

From the points raised in 17.3.1, you will notice that the probability of making type I error is that of rejecting a true null hypothesis, that is, the probability that the test statistic will be in the rejecting region if indeed the null hypothesis is true.

The probability of making a type I error is said to be the *significance level* of the hypothesis test. It is denoted by α. On the other hand, you should note that the probability of making a type II error is that of not rejecting a false null hypothesis.

It is important that you equate this probability to that of the test statistic being in the non-rejection region if indeed the null hypothesis is false. This is denoted by fit and it depends on the true value of the population mean μ.

In making a choice for α, you need to establish a relation between two error probabilities. This is that, for a fixed sample size the smaller the type I error probability, α, of rejecting a true null hypothesis, the larger the type II error probability, β, of not rejecting a false null hypothesis and vice versa.

2.2.4 Possible conclusion for a Hypothesis Test

It is important that you note the following points that

(i) If the null hypothesis is rejected, you can conclude that the alternative hypothesis is true.

(ii) If the hypothesis is not rejected, you can conclude that the data do not provide sufficient evidence to support the alternative hypothesis.

Rejection of the null hypothesis in a test performed at the significant level, α is usually expressed by saying that the test results are *statistically significant* at the α-level.

On the contrary, non-rejection of the null hypothesis in a test performed at the significant level, α is usually expressed by saying that the test results are not *statistically significant* at the α-level.

2.3 Self-Assessment Exercise(s)

The critical concepts learned in this study unit include the following:

1. Statistical inference
2. Hypothesis testing
3. Test statistic
4. Null and alternative hypotheses
5. Rejection and non-rejection region
6. Critical values
7. Level of significance
8. Type I and II errors

It is essential you note the significance tests may be used if it is desired to test whether a population mean (or other numerical summaries) differs from a pre-assigned hypothetical values. This test concerns weighing probabilities and amounts to no proof. It cannot give you information on the origin of disparity beyond reliance on some measures of uncertainty or not.

In nursing and medical practice, however statistical significance cannot be equated with clinical importance.

2.4 Conclusion

In this unit, you have been acquainted with fundamentals of hypothesis testing. You learned the logic behind carrying out a hypothesis test and the precise criterion of drawing conclusions from the test results.

2.5 Summary

1. The Ministry of Health's Food and Drug Agency states that the recommended daily allowance (RDA) of iron for adult females under the age of 50 is 18 mg. a hypothesis test is to be performed to decide whether adult females under the age of 50 are, on the average, getting less than RDA of 18 mg of iron.

(i) Determine the null and alternative hypothesis for the hypothesis test.

(ii) Classify the hypothesis test as two-tailed, left-tailed or right-tailed.

1. Use the information in exercise 17.6.2 to explain what each of the following would mean-

(i) Type I error

(ii) Type II error

(iii) Correct decision

2. Decide whether each of the following statements is true or false. Explain your answer.

(i) If it is important not to reject a true null hypothesis, then, the hypothesis test should be performed at a small significance level.

(ii) For a fixed sample size, a decrease in the significance level of a hypothesis test results in an increase in the probability of making type II error.

2.6 References/Further Readings

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Weiss, A.L (1996) *Elementary Statistics,* 3rd edition, Addison-Wesley Publishing Co. Inc.

UNIT 3 HYPOTHESIS TEST FOR THE POPULATION WHEN STANDARD DEVIATION IS KNOWN

Unit Structure

3.0 Introduction

In this study unit, you will learn a simple method for performing a hypothesis test for a population mean μ, when the standard deviation of the population, σ, is known.

3.1 Intended Learning Outcomes (ILOs)

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

- Carry out a test of hypothesis for a population mean given the standard deviation of the population.

3.2 Main Content

3.2.1 Discussion

You need to be acquainted with how to obtain the critical value(s) for a hypothesis test when the significance level, α, is specified a prior.

You will recall that the significance level, α, of a hypothesis test is the probability of making a Type I error which is the same as the probability of rejecting a true null hypothesis. An equivalent statement is that α is the probability of the test statistic lying in the rejection region if indeed, the null hypothesis is true.

It therefore follows that this is the key to determine the critical values for a specified significance level, which you will learn in what follows.

3.2.1 Obtaining the Critical Value(s) for a Specified Significance Level

You will learn this procedure starting that the following key point:
If a hypothesis test is to be carried out at a significance level, α, then the critical values are chosen such that for a true null hypothesis, the probability equals α so that the test statistic lies in the rejection region.

Let us consider a hypothesis test in terms of this key point for a population mean in the case where the population standard deviation is known. The population is assumed normally distributed or the sample size is taken as large.

Null hypothesis for the test in regard to one population mean, μ is of the form

$H_0 :$ μ = μ_0(μ_0 is a number)

The statistic for the test is

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

The random variable $\overline{x}$ is assumed normally distributed and $\mu_{\overline{x}} = \mu$, the standard deviation $\sigma_{\overline{x}} = \sigma/\sqrt{n}$

If $H_0 :$ μ_0 is true, then the statistic has standard normal distribution. This implies that if the null hypothesis is true, then probabilities for that test statistic are equal to areas under the standard normal curve.

For a specified significance level, α, you need to choose the critical value(s) such that the area under the standard normal curve that lies above the rejection region equal α. The following example will acquaint you with this procedure.

Example 3.1

If a hypothesis test be performed for a population mean, μ, with null hypothesis $H_0 :$ μ = μ_0 known standard deviation and such that the population is either normally distributed or for a large sample size.

Suppose further that the test is performed at the 5% significance level.

Determine the critical value(s) for a (i) two-tailed test (ii) left-tailed test (iii) right-tailed test.

Answer

For α = 0.05, you will choose the critical value(s), with the area under the standard normal curve lying above the rejection region equal to 0.05.

(i) The rejection region for a two-tailed test is on both the left and right. Hence, for a test with α = 0.05, the z-values that divide the area under the standard normal curve into a middle 0.95 area and two outside areas of 0.025 are the required critical values. These are $\pm z_{0.25}$ which we find from the tables to be ± 1.96.

(ii) For a left-tailed test, the rejection region is on the left. Hence, for a test with α = 0.05, $-z_{0.205} = -1.645$ from the tables.

(iii) For a right-tailed test, the rejection region is on the right. Thus, for a test with α = 0.05, $-z_{0.205} = -1.645$ from the tables.

(iv)

These regions are depicted below in Fig 3.1

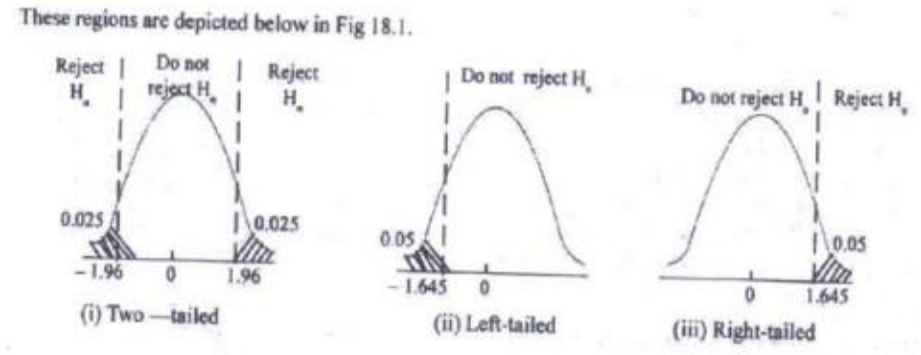


Figure 18.1: *Critical Value(s) of Hypothesis Test*

In summary, if a two-tailed, left-tailed or right-tailed test is respectively performed at a specified significance level, α, the critical values are

$$\pm z_{\frac{\alpha}{2}};\ -z_{\alpha} \text{ and } z_{\alpha}$$

You need to note that the most frequently used significance levels are 0.10, 0.05 and 0.01. you will now learn the procedure for performing a hypothesis test for a population mean when the population standard deviation is specified.

3.2.2 Procedure for a Hypothesis Test for a Population Mean when α is Known

The assumptions are that

(i) Normal population or large sample
(ii) A is specified

Step 1 State the null and alternative hypotheses
Step 2 Specify the significance level α
Step 3 Critical value(s)
(i) Two-tailed test, $\pm z_a$
(ii) Left-tailed test, $-z_a$
(iii) Right-tailed test, z_a
Find values from the Tables of Areas under standard normal curve
Step 4: Calculate the test statistic

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

Step 5: If value of test statistic lies in the rejection region, reject H_0; otherwise, do not reject H_0.
Step 6: State the conclusion in words.
The following example illustrates the procedure outlined.

Example 3.2

A nutritionist thinks the average person with an income below the poverty level gets less than the recommended daily allowance (RDA) 0f 805mg of calcium. To test the conjecture, he obtains the daily intakes of calcium for a random sample of 18 people with incomes below the poverty level. The sample data are given in Table 18.1 below. At the 5% significance level, do the data provide sufficient evidence to conclude that the mean calcium intake of all people with incomes below the poverty level is less than the RDA of 800mg? Assume α = 193mg.

Answer

Sample size *n* = 18 is moderate. A normal probability plot (not shown) for the data reveals no outlier and falls approximately in a straight line.
Step 1: State the null and alternative hypotheses.
Let μ denote the mean calcium intake (per day) of all people with incomes below the poverty level.

H_0 *:* μ = 805mg
H_1 *:* μ < 805mg

(Note that the hypothesis test is left-tailed)
Step 2: describe on the significance level α = 0.05.
Step 3: The critical value for a left-tailed test is $-z_a$
Since α = 0.05, the critical value is $-z_a$. From the standard normal curve tables, $z_{0.05}$ =1.645. so the critical value $-z_{0.05} = -1.645$
Step 4: compute the value of the test statistic

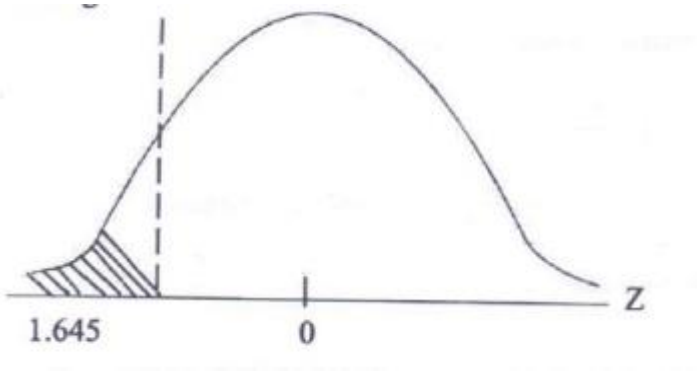
$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{747.4 - 805}{193/\sqrt{18}} = -1.27$$

Step 5: if the value of the test statistic falls in the rejection region, reject H_0.

The value of the test statistic does not fall in the rejection region and so we do not reject H_0.

Step 6: State the Conclusion

The test results are not statistically significant at the 5% level. i.e. at the 5% significant level, the sample of 18 calcium intakes does not provide sufficient evidence to conclude that the mean calcium intake, *p* of all people with incomes below the poverty level is less than the RDA of 800mg.



3.3 Self-Assessment Exercise(s)

The Food and Agency Department of the Ministry of Health gives the recommended daily allowance (RDA) of iron for adult females under 50 as 18mg. the following iron intakes, in milligrams, during a 24-hour period were obtained for 45 randomly selected adult females under the age of 50.

15.0	18.1	14.4	14.6	10.9	18.1	18.2	18.3	15.0
16.0	12.6	16.6	20.7	19.8	11.6	12.8	15.6	11.0
15.3	9.4	19.5	18.3	14.5	16.6	11.5	16.4	12.5
14.6	11.9	12.5	18.6	13.1	12.1	10.7	17.3	12.4
17.0	6.3	16.8	12.5	16.3	14.7	12.7	16.3	11.5

At 1% significant level, do the data suggest that adult females under 50 are on the average, getting less than the RDA of 18mg of iron? You may assume the population standard deviation of 4.2mg (Note that = 14.68mg).

3.4 Conclusion

In this unit, you have learned the simple procedure for performing hypothesis test for a population mean when the population standard deviation is known.

3.5 Summary

- Observe a sample of measurements from the population of interest and compute the sample statistic corresponding to the hypothesized parameter.
- Rejection of a null hypothesis at an α level of significance implying that the investigator is willing to risk an α chance of error in so rejecting H_0.
- The choice of a significance level is arbitrary and depends on the implications of making an incorrect decision. The frequently used values of α are 0.05 and 0.01.
- A Type I error is committed when H_0 is rejected given that H_0 is true.
- Pr(Type 1 error) = α
- A Type II error is committed when H_0 is not rejected when in fact it is false.
- Pr(Type II error) = β
- Procedure for performing an hypothesis test for a population mean, μ, when the population standard deviation, α, is known.

3.6 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses,* 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics,* 12th edition, E.Arnold, London.

Weiss, A.L (1996) *Elementary Statistics,* 3rd edition, Addison-Wesley Publishing Co. Inc.

UNIT 4 CLASSICAL APPROACH VS P-VALUE APPROACH TO HYPOTHESIS TESTING

Unit Structure

4.0 Introduction

This study unit concludes the discussion on hypothesis testing within the scope of this course. It examines the limitations of the classical approach and introduces the *p*-value approach to hypothesis testing. Let us first look at the objectives stated hereunder:

4.1 Intended Learning Outcomes (ILOs)

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

- Apply the *p*-value approach to hypothesis testing.
- Interpret the results of a hypothesis test suitably.

4.2 Main Content

4.2.1 Classical Approach

You will recall the procedure used in performing a hypothesis test for a population standard deviation, α is given. This statistical inference though specific to a particular example, but it is important that you are

aware of the elements common to all hypothesis testing procedures that are bases on the classical approach. These elements are presented as follows:
Step 1: You will state the null and alternative hypotheses
Step 2: You next decide on the significance level, α
Step 3: You will determine the critical value(s).
Step 4: You will next compute the value of the test statistic.
Step 5: If the value of the test statistic falls in the rejection region, then reject H_0.
Step 6: You will finally state the conclusion of the test in words.
Before discussing the limitations of the classical approach to hypothesis testing, you are to examine two other notions related to the subject.

4.2.2 Statistical Significance Versus Practical Significance

You will recall that the results of a hypothesis test are said to be statistically significant if the null hypothesis is rejected at the chosen level of α. The implication of this statement is that the data provided evidence to conclude that the truth is different from that stated in the null hypothesis.

You need to note that this does not necessarily mean that the difference is important in any practical respect. In other words, statistically significance does not imply practical or clinical significance.

However, the decision as to whether a difference is of practical or clinical importance is not a statistical one. Rather the decision lies with those with expertise in the subject area being investigated.

4.2.3 Relation between Hypothesis Test and Confidence Intervals

You need to be aware that there is a close relationship between hypothesis tests and confidence intervals.

If you consider a two-tailed hypothesis test for a population mean at the significance level α, the null hypothesis will be rejected if and only if the value m_o given for the mean in the null hypothesis lies outside the (1-m) = level confidence interval for in.

4.2.4 P-Values

The classical approach to hypothesis testing has the following limitations:

(i) It does not permit readers having access only to the conclusion of the test to make their evaluation (i.e., select their own significance level).
(ii) It does not provide them with the information necessary to access precisely the strength of the evidence against the null hypothesis.

These shortcomings are taken care of by including the p-value of the hypothesis test in the results.

The p-value of a hypothesis test is defined as the probability of observing a value of the test statistic as inconsistent (or more) with the null hypothesis as the value of the test statistic actually observed.
You will therefore notice that the smaller the *p*-value, the stronger the evidence against the null hypothesis. You also need to note that the *p*-value can be referred to as the observed significance level or the probability level.

4.2.5 P-Value Approach to Hypothesis Testing

The elements of the *p*-value approach to hypothesis-testing are as follows:
Step 1: You will state the null and alternative hypotheses
Step 2: You next decide on the significance level, α
Step 3: You will compute the value of the test statistic.
Step 4: You will determine the *p*-value.
Step 5: If P≤ a, then reject H_0; otherwise, do not reject H_0
Step 6: You will state the conclusion of the test in words.
This procedure is illustrated with the following example:

Example 4.1

A dietician believes that an average person with an income below the poverty level gets less than the recommended daily allowance (RDA) of 800mg of calcium. When testing his guess, he obtained the following daily intakes of calcium for a random sample of 10 people with incomes, below the poverty level:

686	994	740	648	1110
992	430	772	670	690

At the 5% significance level, do the data provide sufficient evidence to conclude that the mean calcium intake of all persons with incomes with the poverty level is less than the RDA of 800mh? Assume $s = 180mg$.

Answer

Let μ denote the mean calcium intake (per day) of all people with incomes below the poverty level.

Step 1: H_0: μ = 800mg
H_0 : μ < 800mg
Step 2: The test is to be performed at 5% significance level. So α = 0.05
Step 3: Test statistic is

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

μ_0 = 800, σ = 180, n = 10
from the data,

$$\overline{x} = \frac{\sum x}{n} = \frac{7702}{10} = 770.2$$

$$z = \frac{770.2 - 800}{180/\sqrt{10}} = -29.8$$

$$-0.52$$

Step 4: since the test is left-tailed, the *p*-value is the probability of observing a value of z of -0.524 or less, $P(z \leq -0.524)$, if the null hypothesis is true. That probability equals the shaded area in Fig 19.1 which by the standard normal curve tables is 0.3015.

Figure 19.1: Value of the test statistic and the p-value.
Thus, p = 0.3015.
Step 5: From step 4, p = 0.3015. but this value exceeds the significance level of α = 0.05, you do not reject H_0.
Step 6: The test results are not statistically significant at the 5% level that is at the 5% significance level, the sample of 10 calcium intakes does not provide sufficient evidence to conclude that the mean calcium intake, μ, of all people of incomes below the poverty level is less that the RDA of 800mg.

4.2 Self-Assessment Exercise(s)

A Nursing Council Committee places the mean monthly income of nursing officers as ₦25,000. If a random sample of 100 nursing officers is taken and a mean monthly income of ₦30,000is found and if the population' standard deviation is ₦10,000, at the 1% significance level, do the data provide sufficient evidence to conclude that the mean monthly income of nursing officers is 1425,000?

4.3 Conclusion

In this study unit, you learned a second procedure of hypothesis testing. You also studied the shortcomings of the classical approach, the relationship between confidence intervals and hypothesis tests and the distinction between statistic and practical significance.

4.4 Summary

The critical concepts presented are

(i) P-value of hypothesis test; and

(ii) A difference, although significant may not necessarily be clinically important.

4.5 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses*, 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics*, 12th edition, E.Arnold, London.

Weiss, A.L (1996) *Elementary Statistics*, 3rd edition, Addison-Wesley Publishing Co. Inc.

UNIT 5 MORBIDITY STATISTICS

Unit Structure

5.0 Introduction

This unit involves factors that influence the incidence of illness of different kinds and steps that are likely to contribute to their prevention and cure.

This is possible by means of highly developed vital statistical systems. In the past, the death rate from all causes or from specific causes had been the only significant measure of health progress or lack of it.

In the not distant past, the increase in sickness in sickness not necessarily ending in fatality has called for investigation and preventive measures. In this regard, statistics of sickness present substantial problems. Hence, in medical and nursing practice there is need for the development of a system of morbidity statistics.

5.1 Intended Learning Outcomes (ILOs)

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

- Recognize the need for morbidity statistics
- Identify measures of morbidity
- Set up statistical investigation of morbidity
- Identify sources of morbidity statistics
- Estimate rates of morbidity

5.2 Main Content

5.2.1 Significant Problems of Morbidity Statistics

You need to be aware that statistics of sickness present very substantial problems in comparison to statistics of death. This is because death is a unique event, which occurs at a point in time whereas the occurrence of illness may be repeated in the same person from same or different causes, and its occurrence is for a period of time.

You also realize that death is precisely defined in contrast to illness, which varies in its severity ranging from negligible effects to disabling conditions. Therefore, you need to take these aspects into consideration in the measurements of morbidity.

Its therefore important in your evaluation of the amount of morbidity in a given period that you consider, the following points.

1. Whether you should add the number of persons ill or the illnesses or both to the amount of morbidity.
2. Whether the number of new illnesses that arise in a given period or number that were extent in that period be known.
3. What you intend to count as morbidity in any given circumstances. Is a sickness congenital, acquired effect, injuries, impairment or incipient diseases revealed by test e.g., tuberculosis or diabetes?
4. You need to know about carriers of a disease.

5.2.2 Sources of Morbidity Statistics

You need to know the frequent sources of morbidity statistics and the significant problems arising in each. These are as follows:

1. The survey of sickness with a representative sample of a population keeping a diary or being interviewed so as to reveal the details of the sickness suffered over a defined preceding interval of time.
2. Statistics of a general practitioner of patients attending surgery or visited at home.
3. Hospital in-patient statistics provide a firm diagnosis.
6 Sickness absence records.
7 Notifications of diseases are frequently limited to infectious diseases.
8 Registration of all cases of diseases provides information by which sickness in population may be identified and measured.

It is important that you know that a combination of all these sources of data have meaningful value that contributes to administration research and knowledge.

5.2.3 Rates of Morbidity Statistics

Your decision on appropriate rates of morbidity has to classify the illnesses existing in a population in a time frame as follows:

(i) Illnesses that begin and end during the interval.
(ii) Illnesses that begin during the interval and still exists at the end of the end.
(iii) Illnesses that exist before the beginning of the interval and ending during the interval.
(iv) Illnesses that exist before the beginning of the interval and still existing at the end of the interval.

You will decide for each of these classes on the measure of the number of persons sick or the number of spells of illnesses that occur. The most meaningful morbidity rates in the total population or at specific ages will then turn out to be:

1. The incidence rate, which is the number of illnesses beginning within a specified period of time and related to the average number of persons exposed to risk during that period e.g., how many persons fell sick with typhoid fever in the sixth week of the year?
2. The period prevalence rate which is the number of illnesses existing at any time within a specified period and related to the average number of persons exposed to risk during that period e.g., how many persons were sick with malaria during the month of December?
3. The point prevalence rate which is the number of illnesses existing at a specified period of time and related to the number of persons exposed to risk at the point of time e.g. how many persons were sick with cholera on 24 December.

The average duration of sickness (and the frequency distribution on which it is based).

5.3 Self-Assessment Exercise(s)

State with cogent reasons the need for morbidity statistics.

5.4 Conclusion

In this unit, you learned how to conduct a statistical study of morbidity. To do this, you learned that the crucial aspects of this study include significant problems, sources and rates of morbidity.

5.5 Summary

In this unit, statistics of morbidity was discussed as a supplement of statistics of mortality. This enhances present day principal measure of health progress of a country.

Although the statistics of morbidity present very substantial problems, yet in research work or statistical survey, the method of collection of data and details of analysis need be supplied.

The most usual measures of morbidity are the incidence and prevalence rates.

5.6 References/Further Readings

Knapp, R.G (1985). *Basic Statistics for nurses*, 2nd edition, Delmar Publishers Inc. N.Y.

Hill, A.B. and Hill, I.D. (1991). *Principles of Medical Statistics*, 12th edition, E.Arnold, London.

Weiss, A.L (1996) *Elementary Statistics*, 3rd edition, Addison-Wesley Publishing Co. Inc.